Exploring the genetic landscape of complex diseases using the recessive model

Exploring the genetic landscape of complex diseases using the recessive model

A dissertation presented

by

Teng Ting Lim

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Genetics and Genomics

Harvard University

Cambridge, Massachusetts

April, 2014

Dissertation Advisor: Professor Mark J. Daly Teng Ting Lim

Exploring the genetic landscape of complex diseases using the recessive model

ABSTRACT

High-throughput sequencing technologies have changed the way we identify, study and understand the role of rare variation in Mendelian diseases. Sequencing in complex diseases have proven to be more challenging to interpret, but methods and approaches are being developed to aid in our understanding of variation in these diseases.

In this dissertation, we have sought to interpret and understand the role of rare (<1% allele frequency) and low-frequency (1-5% allele frequency) variants in the genetic etiologies of complex diseases such as autism. We compared the rates of rare 2-hit (homozygous and compound heterozygous) loss-of-function variants in ~1,000 autism cases and ~1,000 controls and discovered an excess of such events in the cases, suggesting that ~5% of cases might harbor a rare 2-hit loss-of-function variant that confers risk for their disorder.

Next, we developed a novel statistical method (RAFT) and discovered 3 individuals with autism and intellectual disability who harbor rare homozygous missense mutations in the cholesterol biosynthesis gene *DHCR24*. We adapted a yeast biochemical assay to understand the efficiency of cholesterol synthesis for these missense variants in *DHCR24*, as well as a population survey of all missense variants from 4,300 European Americans.

Finally, we utilized the unique genetic architecture as a result of bottlenecks by demonstrating that such populations are enriched for rare deleterious variants that might have

important medical consequences. By genotyping 83 loss-of-function variants in 36,000 Finns, we discovered several associations, including a strong protective association between *LPA* and coronary heart disease, suggesting that knocking out *LPA* might prove to be an effective drug target in humans.

TABLE OF CONTENTS

ACKNOWLEDGEMENT

First and foremost, I would like to thank my Ph.D. advisor Mark Daly for his mentorship and guidance on my projects. I have learnt a lot from you in terms of intuition for human genetics data, as well as statistical and population-based approaches to solving problems. More importantly, I appreciate your patience and support in these difficult projects that required a long time to mature and involved working with lots of different groups. I first learnt statistical genetics from reading a C program that you wrote in 1997 and had spent several months asking around trying to decipher the first version of RECMOD that you wrote in 20 minutes. It was funny how unconventional my learning experience had been, but somehow it worked.

Secondly, I would like to thank my "academic big brother" Soumya Raychaudhuri who has been a pillar of support and encouragement all these years. I had lots of fun brainstorming, exploring ideas and performing simulations together for our first project. I value your encouragement, mentorship and advice for everything in general, ranging from scientific to non-scientific issues. I recall that for a really long time, I didn't really want to do some power calculations that you had suggested since they were a bit of a hassle, but I finally caved in and eventually did all the calculations. Thanks for being supportive all these years and for lending a helping hand when things got tiring and tough! ☺

I would like to thank all past and present members of the Daly/Altshuler labs. It has been great to be able to seek help from the huge number of supportive post-docs and grad students in the lab. I enjoyed the lab meetings and critical feedback from David Altshuler all these years. I would also like to specially thank the Hirschhorn lab (including Joel Hirschhorn) for inviting me along to meetings, outings and coffee breaks, and more importantly, for treating me like a

member of the lab. I remembered walking into the lab one day announcing that I wanted to learn quantitative trait analyses so that I could scoop you guys and work on height/BMI as well. Everyone laughed and started sharing the Hirschhorn trade secrets and background information for performing quantitative trait analyses, including Joel himself. I really appreciate that.

I would also like to thank my dissertation advisory committee (DAC) George Church, Jim Gusella and Steve McCarroll. I appreciate the advice and feedback all these years, even when my projects were not working or when I had nothing to present during my DAC meetings except some potential project ideas. I am also extremely grateful for the informal and causal mentor-mentee relationships with every one of you outside the DAC meetings. It has been wonderful to know that I could always get advice or just blabber updates on my projects every time I bump into you guys at NRB, Starbucks or some other random places.

I would like to thank my unofficial academic advisor Ting Wu, who took on the role after my official academic advisor Fritz Roth left Harvard. I enjoyed the times when I was rotating with Steve McCarroll and would walk into your office every now and then to learn about structural variations. I liked how you think deeply about the field and shared all these knowledge and insights, which fascinated me a lot. And you have continued to be a mentor for me whenever I was feeling lost and needed advice for anything in general.

My parents and siblings have been a wonderful source of support during my graduate school. We skype every Sunday and it felt that I was still in Singapore, giving you guys regular updates on life here. There were times when graduate school wasn't all that rosy and I appreciate the concern and support and you guys coming over here for a visit. I hope that you are all proud of me and will continue to support my interest in science.

I thank my good friends from graduate school Palak Amin, Laura Liu and Ying Kai for the weekly weekend dinners and games nights. I have significantly improved in playing video, board and card games over the past few years. We have so much fun together and can read one another so well that it has been difficult to play Bridge together without giving away anything. A raised eyebrow, a slight smile and even a small pause are signs that we have learnt to read from one another too well. I hope to continue to hang out with you guys for our regular weekend dinners and games nights, and I hope that you guys will take up my suggestion and wear paper bags when playing Bridge.

Last but most importantly, I thank my soul-mate, my best friend, my best collaborator and my husband Rigel Yingleong Chan. Rig, you have always been there with me through happy and depressing times. You have even helped me with ideas and analyses for my projects whenever I was stuck. When I was feeling lost and frustrated, you had asked your Ph.D. advisor Joel if he could help provide advice and suggestions for my research projects. I was extremely touched by the thought and gesture. More importantly, I appreciate you doing all the cooking as I absolutely detest eating the food that I've cooked. I dedicate my dissertation to you and look forward to exploring new frontiers with you soon!

ATTRIBUTIONS

Chapter 2

Elaine T. Lim: Conceived of and designed the experiments, performed analyses, wrote manuscript and edited later drafts based on comments from advisors (M.J.D. and S.R.), co-authors and reviewers.

Soumya Raychaudhuri: Contributed to the conception and design of the experiments, provided statistical guidance for the analyses, structured the results and edited the manuscript and response letters to the reviewers.

Stephan J. Sanders and Matthew W. State: Provided replication data, as well as helpful edits and comments for the manuscript.

Christine Stevens, Aniko Sabo, Benjamin M. Neale, Uma Nagaswamy, Donna Muzny, Jeffrey G. Reid, Alicia Hawes, Irene Newsham, Yuanqing Wu, Lora Lewis, Huyen Dinh, Shannon Gross, Li-San Wang, Chiao-Feng Lin, Otto Valladares, Stacey B. Gabriel, Mark dePristo, NHLBI Exome Sequencing Project: Provided data for initial discovery and genotyping validation.

Daniel G. MacArthur and Monkol Lek: Provided advice for annotating the loss-of-function variants, as well as comments for the manuscript.

Andrew Kirby: Provided advice and suggestions for interpreting the results, as well as edits and comments for the manuscript since the first draft.

Douglas M. Ruderfer, Menachem Fromer and Shaun M. Purcell: Replicated the enrichment analyses on the autism exome sequencing data, provided copy number data and helpful edits and comments for the manuscript.

Li Liu, Kathryn Roeder and Bernie Devlin: Provided coverage information on the exome sequencing data for the supplementary material, as well as critical edits and comments for the manuscript.

Jason Flannick and David M. Altshuler: Provided advice and technical help in analyzing whole-exome sequencing data, as well as critical edits and comments for the manuscript.

Eric Boerwinkle, Joseph D. Buxbaum, Edwin H. Cook, Richard A. Gibbs, Gerard D. Schellenberg and James S. Sutcliffe: Provided data for initial discovery and critical comments and feedback for the work and manuscript.

Mark J. Daly: Provided discovery and replication data, contributed to the conception and design of the experiments, provided statistical guidance for the analyses, structured the results and edited the manuscript and response letters to the reviewers.

Chapter 3

Elaine T. Lim: Conceived of and designed the experiments, performed analyses and experiments, wrote manuscript with comments from advisors (M.J.D. and T.W.Y.).

Yingleong Chan, Joel N. Hirschhorn and Soumya Raychaudhuri: Performed simulations on Finns and non-Finnish Europeans to help characterize the statistical method we have been developing, as well as critical feedback and advice for working out issues in the method and critical edits and comments for the manuscript.

Susanne Goetze: Helped with bacterial transformations, minipreps and maxipreps.

Daniel Spatt and Federick Winston: Provided reagents, critical help and feedback on the yeast experiments.

Lisa Kratz and Richard I. Kelley: Performed mass spectrometry experiments to detect sterol levels from human serum and yeast lysates, as well as advice for experiments.

Phil Lee, Jacqueline I. Goldstein and Christine Stevens: Provided exome chip data and performed quality control on the data.

Douglas M. Ruderfer, Shaun M. Purcell: Implemented RAFT in the software package PLINK/SEQ, as well as provided advice on the statistical method.

Maria Chahrour, Matthew Johnson, Chris A. Walsh: Provided Middle Eastern exome sequences and feedback on the project.

Mark J. Daly: Provided statistical guidance for the development of the recessive test, as well as comments and edits for the manuscript.

Timothy W. Yu: Provided experimental guidance, as well as provided comments and edits for the manuscript.

Chapter 4

Elaine T. Lim: Conceived of and designed the experiments, performed analyses, wrote manuscript and edited later drafts based on comments from advisors (M.J.D. and A.P.), co-authors and reviewers.

Peter Würtz: Provided phenotype data, replicated analyses, performed Cox regression analyses on the *LPA*-MI results and provided helpful edits for the manuscript.

Aki S. Havulinna: Provided phenotype data and helpful edits for the manuscript.

Pritt Palta: Performed quality control and re-calling of genotypes from the Sequenom experiment based on the cluster plots.

Taru Tukiainen and Karola Rehnström: Performed the initial genotype quality control and association analyses.

Tõnu Esko, Reedik Mägi and Andres Metspalu: Provided replication data for the *LPA*-MI association from the Estonian Biobank.

Michael Inouye: Provided microarray gene expression data and analyses for the *MS4A2*-TG results, as well as helpful feedback for the manuscript.

Tuuli Lappalainen: Provided RNA sequencing data and analyses for the Mendelian genes, as well as helpful feedback for the manuscript.

Xueling Sim, Alisa Manning, Claes Ladenvall and Cecilia M. Lindgren: Provided replication data for *MS4A2*-TG and *ATP2C2*-SBP associations.

Suzannah Bumpstead and Eija Hämäläinen: Performed the Sequenom genotyping experiments and quality control at Sanger Institute.

Kristiina Aalto, Mikael Maksimow, Marko Salmi, Jeffrey C. Barrett, Terho Lehtimäki, Markku Laakso, Leif Groop, Jaakko Kaprio, Markus Perola, Mark I. McCarthy, Michael Boehnke, David M. Altshuler, Nelson B. Freimer, Tanja Zeller, Sirpa Jalkanen, Seppo Koskinen, Olli Raitakari and Richard Durbin: Provided data for intial discovery and replication, as well as critical comments for the manuscript.

Stefan Blankenberg, Diego Ardissino, Svati Shah, Benjamin Horne, Ruth McPherson, Gerald K. Hovingh, Muredach P. Reilly, Hugh Watkins, Anuj Goel, Martin Farrall, Domenico Girelli, Alex P. Reiner, Nathan O. Stitziel, Sekar Kathiresan and Stacey Gabriel: Provided replication data for the *LPA*-MI association.

Joel N. Hirschhorn: Provided critical feedback for the project and advice for structuring the results for an oral presentation, as well as critical comments for the manuscript and reviews.

Daniel G. MacArthur, Veikko Salomaa and Samuli Ripatti: Involved in the design of the experiments and provided critical comments for the manuscript and reviews.

Mark J. Daly and Aarno Palotie: Conceived the experiments, directed research and analyses, as well as edited the manuscript and response letters to the reviewers.

CHAPTER 1

Introduction

A PREAMBLE

Genetic mapping in the 1980s

In 1983, James Gusella and his colleagues first discovered a polymorphic DNA marker on 4p16.3 associated with Huntington's disease through the use of a then-new technology called restriction fragment length polymorphism (RFLP) genotyping and linkage analysis in 2 families [1]. It took nearly a decade of research before scientists identified the gene (Huntingtin; *HTT*) and the trinucleotide repeat $(CAG)_n$ in *HTT* cause the disease [2]. Shortly thereafter, a series of genetic studies into various diseases using similar technologies soon discovered the mutations and causal genes involved in human diseases, such as ΔF508 in the Cystic Fibrosis Transmembrane Conductance Regulator (*CFTR*) gene involved in cystic fibrosis [3]. The statistical evaluation for these linkage studies were laid decades earlier when Newton Morton derived a score for calculating the logarithm of the odds, or the LOD score for evaluating the significance of genetic discoveries from linkage studies in pedigrees in 1955 [4].

The invention of new technologies is often quickly accompanied by a wave of new methods and approaches for genetic mapping and statistical evaluation of genetic discoveries. By 1987, in anticipation of a complete RFLP linkage map of the human genome, Eric Lander and David Botstein introduced "homozygosity mapping" for mapping recessive genes in inbred families [5]. The concept behind the approach was to calculate the probability that an affected child with inbreeding coefficient F has homozygosity by descent at a disease allele with frequency q is $\frac{Fq}{Fq+(1-F)q^2}$. Subsequent papers by Eric Lander, Nicholas Schork and Leonid Kruglyak demonstrated that a LOD score of 3.3 instead of 3 is required to achieve a genome-wide false positive rate of 5% [6,7]. However, the likelihood computations in traditional linkage calculations increase exponentially with the number of loci and haplotypes, so methods such as

the Elston–Stewart and Lander-Green algorithms were devised to calculating the likelihoods in pedigrees [8,9]. In 1995, a model for calculating recessive LOD score was also proposed and implemented [10]. To-date, tools such as homozygosity mapping, LOD score calculations and the genome-wide significance LOD score threshold of 3.3 have been widely adopted by human geneticists for mapping genes underlying recessive diseases in consanguineous populations [11,12,13].

Genetic mapping in the early 21th century

In 2005, shortly after the completion of the Human Genome Project, the HapMap Project was completed with several groups involved in characterizing naturally occurring human variation in diverse populations [14]. This resulted in the development and use of single nucleotide polymorphism (SNP) genotyping arrays, as well as copy number variant (CNV) genotyping arrays. As the development of the arrays proceeded with increasing density of the markers, as well as figuring out the optimal lengths of the oligonucleotides, the development of tools and methods for the discovery of variants associated with various diseases proceeded as well. One of the earliest genome-wide studies was a “transmission test for linkage disequilibrium” (TDT) performed using 290 autosomal microsatellite markers for a Mendelian disease Saguenay-Lac-Saint-Jean cytochrome oxidase deficiency and the authors discovered an genome-wide linkage disequilibrium region on chromosome 2p16 with a p-value of 1.2×10^{-5} [15]. The first successful genome-wide association study (GWAS) in a complex disease was performed in 2005 using a chi-squared test on the allele frequencies of 116,204 SNPs in 96 European cases with age-related macular degeneration and 50 controls [16]. The authors

discovered 2 significantly-associated SNPs in the intron of the complement factor H (*CFH*) gene that surpassed the study-wide significance threshold of 4.8×10^{-7}.

Several methodologies and standards such as the use of genomic control [17] to evaluate spurious associations driven by population stratification between cases and controls were developed and adopted for these GWAS. To-date, several case-control association studies have been performed for various common diseases and traits such as height [18]. An alternative approach was developed in 1993 by Spielman *et al.* called the TDT (as mentioned earlier) [19]. In short, TDT counts the number of transmitted and untransmitted alleles from heterozygous unaffected parents to heterozygous affected children and uses the McNemar's chi-squared test statistic to evaluate probability. One advantage in performing TDT as compared to the case-control GWAS tests is that TDT is not affected by population stratification – that is, systematic ancestry differences between cases and controls that can confound case-control association studies.

An early GWAS performed in 2009 for schizophrenia with 3,322 cases and 3,587 controls did not discover any loci with genome-wide significance [20], but a subsequent GWAS study in 2013 with 8,832 schizophrenia cases and 12,067 controls discovered 13 significant loci [21]. While GWAS using SNP genotyping arrays has been successful for several complex diseases and traits such as height where hundreds of loci have been discovered to contribute to the phenotype [18], GWAS has not yet worked for some complex disorders such as autism. To-date, no single significant locus from GWAS in autism has been successfully replicated across different studies [22,23,24]. However, a recent study has estimated the contribution of common variants in autism to be ~40% for simplex families and ~60% for multiplex families [25],

suggesting that common variants associated with autism should be discovered with larger sample sizes.

However, the widespread ease and use of the CNV genotyping arrays led to a series of groundbreaking discoveries on the importance of recurrent *de novo* deletions and duplications associated with autism [26]. In 2009, Jonathan Sebat and Michael Wigler performed a genome-wide scan on 85,000 probes in 264 families and observed a 10-fold excess of *de novo* CNVs in individuals with sporadic autism (that is, there is only 1 affected individual in the family), and a 3-fold excess of *de novo* CNVs in affected individuals with another affected first-degree relative. While the authors could not identify any specific CNV associated with risk for autism, they discovered the importance of *de novo* CNVs with autism ($P = 0.0005$). Recent studies using microarrays with more probes and larger sample sizes on >500 families have implicated specific *de novo* loci such as 16p11.2 with autism [27,28,29]. A subsequent study using CNV genotyping arrays on 104 Middle Eastern families has also implicated the role of rare recessive copy number variants with autism, although no specific locus could be identified with confidence [13]. While the role of *de novo* and rare inherited CNVs have been associated with increased risk for autism, some of these CNVs span across regions as large as several megabases long and across tens of genes. As such, it has been difficult to pinpoint specific genes underlying some of these CNVs that are associated with autism.

Genetic mapping in post-2009 for Mendelian diseases

More recently, in the era of post-GWAS performed using SNP and CNV arrays, human geneticists have recently adopted a range of methodologies for studying disease genetics through the use of whole-genome sequencing or whole-exome sequencing, which has revolutionized the

field of human genetics, especially for the discovery of rare variants with large effects. As an example, we revisit a landmark paper in 2009 when Ng *et al*. sequenced 8 control individuals from HapMap and 4 individuals with a dominant disease called Freeman-Sheldon syndrome, caused by mutations in *MYH3* and as a proof-of-concept, demonstrated that they were able to re-discover *MYH3* as the causal gene using whole-exome sequencing [30]. Subsequently, more causal genes in various Mendelian diseases were identified [31,32]. To-date, whole-exome sequencing has been used routinely to identify novel genes that are involved in rare Mendelian diseases [33,34,35,36]. A pilot study to evaluate the utility of using whole-exome sequencing to diagnose 250 probands with a range of neuro-developmental phenotypes discovered that approximately 25% of the probands were able to obtain a genetic diagnosis [37], highlighting the potential importance of exome sequencing in the near future for clinical diagnosis of rare Mendelian diseases. We have been involved in helping clinical collaborators with mapping novel genes and variants for various recessive Mendelian diseases such as neuronal ceroid lipofuscinosis, where we found a novel causal gene in the potassium channel tetramerization domain-containing protein 7 (*KCTD7*) gene [34], diacylglycerol acyl transferase 1 (*DGAT1*) in infantile enteropathy [33], as well as digenic mutations in *OTUD4* and *RNF216* in ataxia and hypogonadism [35].

GENETIC MAPPING IN COMPLEX DISEASES AND TRAITS

Initial studies using whole-exome sequencing in complex diseases and traits such as combined hypolipidemia, adopted similar screening and filtering strategies as those used for Mendelian diseases. This was motivated by the assumption that complex diseases might be driven by several rare variants and genes of large effects, similar to the causal variants and genes

found in Mendelian diseases. This assumption was supported by earlier work that discovered variants of large effects in complex diseases and traits, such as the proprotein convertase subtilisin/kexin type 9 (*PCSK9*) gene. A linkage study performed in 2003 on a large pedigree with 35 individuals discovered a missense mutation (S127R) in *PCSK9* that was associated with hypercholesterolemia and increased risk for coronary heart disease [38]. In 2005, Cohen *et al*. tested the hypothesis if loss-of-function mutations in *PCSK9* have a similar or opposite effect as the S127R mutation and sequenced the coding regions of the gene in 198 African-Americans, and discovered 2 nonsense mutations (Y142X and C679X) that were associated with reduced levels of circulating low-density lipoprotein (LDL), thus potentially reducing risk for coronary heart disease [39]. This was subsequently proven in a later study [40].

Similarly, in 2009, Musunuru *et al*. performed whole-exome sequencing on 2 probands from a large pedigree of 38 individuals with combined hypolipidemia, characterized by low levels of low-density lipoprotein (LDL), high-density lipoprotein (HDL) and triglycerides, and discovered compound heterozygous loss-of-function variants in the angiopoietin-like 3 protein (*ANGPTL3*), suggesting that similar to the loss-of-function variants in *PCSK9*, loss-of-function variants in *ANGPTL3* might potentially reduce risk for coronary heart disease as well [41]. However, such studies proved to be extremely rare success stories in complex disease genetics. In reality, whole-exome sequencing has proven to be more challenging for discovering novel or causal genes and variants underlying risk for complex diseases such as autism and schizophrenia.

Whole-exome sequencing to discover rare polygenic and de novo variants

A recent study with whole-exome sequencing of ~2,500 schizophrenia cases and ~2,500 controls conducted by Purcell *et al*. demonstrated that while there are certain sets of genes such

as activity-regulated cytoskeleton-associated scaffold protein and targets of the fragile X mental retardation protein (FMRP) that were enriched for mutations in the cases, they could not detect evidence for specific genes or rare alleles (0.5-1% allele frequencies) with large effects [42]. In a recent 2014 study by Yingleong Chan and Joel Hirschhorn, they developed a novel method for detecting rare polygenic signals from summary results obtained in GWAS of various common diseases [43]. The approach utilizes the observation that rare risk variants are easier to detect than rare protective variants in a case-control study. Using their approach, Chan *et al.* demonstrated that there are strong rare (<1% allele frequency) and low-frequency (1-5% allele frequency) polygenic signals involved in conferring risk for schizophrenia, consistent with the observation that schizophrenia is highly polygenic and that there are rare and low-frequency variants to be discovered in the future with larger sequencing studies.

Similarly, a study involving whole-exome sequencing of ~1,000 autism cases and ~1,000 controls did not uncover evidence for specific genes or variants involved in autism [44]. And subsequently, a larger study by Lee *et al.* with almost 5,000 autism cases and more than 10,000 parental and population controls using exome chip genotyping, which involves genotyping a subset of variants discovered from whole-exome sequencing, did not uncover evidence for specific genes or variants either (unpublished). However, using the same method developed by Chan *et al.*, the authors found a strong rare and low-frequency polygenic signal in autism, again suggesting that larger sample sizes are needed for gene and variant discovery. In epilepsy genetics, whole-exome sequencing in ~900 cases and ~2,000 controls did not identify rare variants of large effects as well [45].

On the other hand, an initial study in 2011 explored the role of *de novo* point mutations from whole-exome sequencing data on 20 trios with autism and discovered 4 potentially

interesting genes (*GRIN2B*, *LAMC3*, *FOXP1*, *SCN1A*) [46]. Subsequent studies with almost 1,000 trios discovered a strong contribution from *de novo* loss-of-function variants in autism and provided evidence for specific genes such as *SCN2A*, *CHD8*, *CTNNB1* and *DYRK1A* involved in autism risk [47,48,49,50]. Whole-exome sequencing to identify *de novo* point mutations involved in intellectual disability and epilepsy have also proven to be fruitful and have implicated several genes such as *GABRB3*, *CACNA1A*, *CHD2*, *FLNA*, *GABRA1*, *GRIN1*, *GRIN2B*, *HNRNPU*, *IQSEC2*, *MTOR* and *NEDD4L* [51,52]. On the other hand, whole-exome sequencing studies to identify *de novo* point mutations have not successfully identified specific genes in other complex diseases such as schizophrenia [53]. However, it is worth noting that for intellectual disability, autism and epilepsy, clear excesses of *de novo* point mutations have been identified in the cases compared to unaffected siblings or controls, whereas such an excess has not been observed for schizophrenia.

AUTISM GENETICS

Using autism genetics as an example, previous literature has demonstrated that *de novo* CNVs have been shown to contribute an estimated 10-15% risk to autism and specific loci such as 16p11.2 have been robustly associated with autism risk [28,29]. Homozygous CNVs have also been shown to play a role in autism [13], although Middle Eastern consanguineous families were used in the study and it is unclear how much these homozygous CNVs could account for in an out-bred European population. On the other hand, common SNPs have been estimated to contribute 40% risk to autism, although no single locus has been robustly discovered [25]. Rare polygenic inheritance have also been discovered to contribute to autism risk (unpublished), although no specific locus can be implicated and it is unclear what the contribution of such

variation is to autism. More recently, *de novo* point mutations discovered from whole-exome sequencing have been estimated to contribute 20-50% risk to autism and specific recurrent genes have been discovered [47,48,49,50].

Addressing heritability and gender bias in autism

However, as we have learnt from Mendelian genetics, there are several modes of inheritance involved in disease etiologies, such as the *de novo* or dominant, recessive, as well as X-linked dominant and recessive models. These studies to identify *de novo* mutations in autism do not address a fundamental question about the high heritability of autism ($h^2 \sim 0.8$) and that there are many families with multiple affected children. In addition, a second interesting observation is that there are typically 4 affected males for every affected female with autism, and this observation cannot be addressed by the *de novo* model in autism as such *de novo* point mutations and CNVs have been shown to occur in both genders equally [28,29,47,48,49,50]. In order to explore the role of inherited variation in autism, we tested the hypothesis if rare recessive variants have a significant role in autism (Chapter 2) and discovered that there is a ~2-fold excess of rare (≤5% allele frequency) homozygous and compound heterozygous loss-of-function variants on the autosomes in probands with autism and that this excess was present in genes found to be expressed in the brain. A similar excess was observed for rare (≤0.25%) hemizygous loss-of-function variants outside the pseudo-autosomal regions on the X-chromosome in affected males compared to unaffected males. Collectively, we estimated a significant 5% contribution to autism from these "rare complete knockouts" of genes.

Many of these genes were observed only once and we could not implicate specific genes in this study, similar to some of the earlier studies implicating the role of *de novo* CNVs and

point mutations. However, some of the candidate genes were reported as known disease-causing genes such as *USH2A* which causes deafness and blindness. We confirmed with the clinicians that the proband with compound heterozygous loss-of-function variants in *USH2A* had severe hearing problems, consistent with the phenotypic description for Usher Syndrome. The other genes were Fragile X E mental retardation syndrome protein (*AFF2*), *KIAA2022* which was previously implicated in intellectual disability, Sushi-repeat containing protein, X-linked 2 (*SRPX2*) involved in rolandic epilepsy and methyl CpG binding protein 2 (*MECP2*) involved in Rett Syndrome. It is worth noting that hemizygous loss-of-function mutations in *MECP2* typically result in death among males, so only females with Rett Syndrome are viable. However, the mutation we discovered in *MECP2* truncates only the last 4 amino acids of the protein and is found to be heterozygous in the mother and hemizygous in the proband and his affected brother, potentially suggesting that this late-truncating mutation results in a largely functional protein and is viable in males.

Our discovery of the autistic proband with compound heterozygous loss-of-function variants in *USH2A* also raised an interesting question regarding incidental discoveries through the use of whole-exome sequencing in research studies. The protocols and avenues for returning such findings from incidental discovery have yet to be defined and established. While the American College of Medical Genetics and Genomics (ACMG) has recently proposed a set of guidelines and genes for returning medically-actionable information to the patients [54], we note that *USH2A* is not among the list of genes listed by ACMG. However, given that the proband with the compound heterozygous loss-of-function variants in *USH2A* did have severe hearing loss, we reported the result back to the clinic that the patient had been enrolled at. Thereafter, the variants were confirmed in a CLIA lab and IRB was approached about returning information to

the primary care physician, and ultimately the patient, as the original consent for our research study had not included re-contacting or return of information to the participants. The finding was deemed actionable and ultimately, genetic counseling was provided to the family with respect to this genetic diagnosis.

Other literature supporting the recessive model in autism

During the time when we were working on the manuscript for Chapter 2 and subsequently, other studies have also discovered an excess of runs of homozygosity in autism cases, further supporting a role for the recessive model in the disorder [55,56]. A pair of papers discovered initial evidence for specific genes that are potentially involved in conferring risk to autism in an autosomal recessive mode of inheritance, such as *UBE3B*, *SYNE1* and *AMT* [11,12]. More recently, a paper provided conclusive evidence from 3 Middle Eastern families for the role of homozygous deleterious mutations in the branched chain ketoacid dehydrogenase kinase (*BCKDK*) with autism and epilepsy [57]. This work was of potential therapeutic interest as well, as the authors demonstrated that they were able to reverse the phenotype in mice with dietary supplementation. This provided evidence that a subset of individuals affected with autism can potentially benefit from a dietary treatment. However, recessive mutations in *BCKDK* are extremely rare outside the Middle Eastern families and no European individual with recessive mutations in *BCKDK* have been discovered thus far and this raises the question of whether such treatment can be generalized to help affected children outside the Middle Eastern populations.

The recessive model in other complex diseases

In collaboration with groups studying Type 2 diabetes and schizophrenia, we have also applied a similar approach to evaluate the contribution of the recessive model in these complex diseases. However, we did not observe any excess of rare or low-frequency recessive loss-of-function variants in Type 2 diabetic cases or schizophrenia cases, even with larger sample sizes than our autism study (unpublished). This suggests that the recessive loss-of-function variants might not play a huge role in the genetic etiologies of such diseases, but these analyses do not rule out the possibility for some recessive variants or genes to confer risk in Type 2 diabetes or schizophrenia.

Novel statistical test to identify rare recessive variants and genes

In Chapter 3, we first developed a novel statistical test for identifying rare recessive variants in complex diseases. Power calculations showed that conventional statistical tests such as logistic regression or Fisher's Exact Test on the homozygous counts were underpowered for detecting rare recessive variants <5% allele frequency. As such, we developed a test that increases the power for detection by utilizing the deviation from the expected probabilities of the homozygotes. The intuition behind the approach is that an observation of 5 cases that are homozygous for a rare variant with 0.1% allele frequency should be more unusual than an observation of 5 cases that are homozygous for a common variant with 10% allele frequency. We termed our statistical test as RAFT (for Recessive Allele Frequency-based Test).

However, one issue with the RAFT test is that deviation from the expected probabilities is also indicative of genotyping errors (such as hybridization issues on SNP genotyping arrays) or misalignment errors from sequencing data, resulting in erroneous variant calls. As such,

stringent filtering and quality control steps are required and we have tested and developed a series of methodologies for identifying and removing such false positives. In addition, if the populations used in the study are of non-homogeneous populations (such as Finns and non-Finnish Europeans), it can result in an excess of homozygosity for both rare and common variants. To address this, we have also formulated a series of analytic strategies for identifying such scenarios.

Discovery of DHCR24 in autism

We applied RAFT to an exome chip genotyping dataset comprising of ~1,000 unrelated probands with autism of European ancestry and ~2,000 unaffected parents as controls. In doing so, we discovered that the top hit was a rare (0.05% allele frequency) missense variant in the 24-Dehydrocholesterol Reductase (*DHCR24*) gene that was highly conserved in 46 vertebrates and was predicted by PolyPhen2 to be deleterious, and that this rare missense variant segregated in an autosomal recessive manner in a family with 3 affected children. Subsequently, we discovered another 2 homozygous private missense variants in 2 Middle Eastern families with autism and intellectual disability. *DHCR24* is a key enzyme involved in converting desmosterol into cholesterol in the liver and brain. In order to assess the pathogenicity of the missense variants, we adapted a yeast biochemical assay developed by Waterham *et al*. [58] and characterized the conversion rate of desmosterol to cholesterol for the 3 missense variants discovered from these families.

These missense variants in *DHCR24* were likely to result in loss-of-function for the gene and that we found fetal demises harboring nonsense and deleterious missense mutations, suggesting that knocking out both copies of *DHCR24* is not tolerated in healthy humans.

However, it has been shown for a similar disorder Smith-Lemli-Opitz syndrome (SLOS) marked by low levels of cholesterol as a result of recessive mutations in the *DHCR7* gene, that simvastatin or cholesterol supplementation such as egg yolks might aid in improving aberrant behavior [59]. However, a larger study has questioned the role of cholesterol supplementation in the treatment of individuals affected with SLOS, given that cholesterol supplementation does not typically cross the blood-brain barrier and that cholesterol synthesis in the brain occurs *de novo* [60].

Potentially treatable subsets of complex diseases

During the process of our work on the *DHCR24* discovery, a notable paper on a suppressor screen in Rett Syndrome was published and highlighted a nonsense mutation in the Squalene monooxygenase (*SQLE*) gene that suppressed the Rett Syndrome phenotype in *MECP2*-null nice, thus presenting cholesterol-lowering drugs such as simvastatin as a potential treatment for individuals with Rett Syndrome [61]. Another important paper by Novarino *et al.* that discovered a potentially treatable subset of autism with recessive mutations in the *BCKDK* gene demonstrated that there might be a small number of individuals with deficiencies in key metabolic genes that result in severe to mild neuro-developmental disorders [57]. A third paper followed with Ruzzo *et al.* describing a rare form of intellectual disability caused by recessive mutations in the asparagine synthetase (*ASNS*) gene which catalyzes the synthesis of asparagine from glutamine and aspartate, suggesting that asparagine supplementation can potentially aid these affected individuals if treatment was performed early [62]. Together with our work on *DHCR24*, while these papers discovered potentially treatable subsets that account for only a small percentage of all individuals affected with autism and other neuro-developmental

disorders, they present a new and exciting research focus by utilizing new genomics technologies (such as whole-exome sequencing) for the rapid discovery of genetic causes underlying complex diseases, functional validation of the genes and variants discovered and potential treatment options for these affected individuals.

However, it should be cautioned that the path towards treatment or therapeutics even for such potentially treatable subsets might require several more decades of research. While it has been shown that cholesterol supplementation for children affected with SLOS might help to improve aberrant behavior, it has not been demonstrated convincingly that post-natal supplementation can help in improving the intellectual and cognitive abilities of these children. It will be important to assess and understand the impact of cholesterol deficiency temporally throughout the development of the human brain. One potential solution is for pregnant mothers who have fetuses that are at risk for cholesterol deficiency diseases might be placed on high cholesterol diet, although long-term assessment of such a diet on the mothers' health will be required. Moreover, since cholesterol is produced in the brain *de novo*, it is unclear if cholesterol supplementation can aid in increasing the levels of cholesterol in the developing brain.

FINNISH GENETICS

Traditionally, the genetic causes for several recessive Mendelian diseases were mapped in founder populations such as Ashkenazi Jewish and Finnish individuals [63,64,65] or consanguineous inbred populations such as the Middle Eastern individuals [66,67,68]. To further understand the role of recessive genetic factors in common diseases, we explored a unique founder population such as the Finns. We demonstrated in Chapter 4 that a variant that causes embryonic lethality in a dominant mode of inheritance will be virtually absent or can found at

extremely low allele frequencies in both the Finnish and out-bred European populations. However, this is not true for a recessive variant that results in embryonic lethality. Even in the worst case scenario where a recessive variant results in embryonic lethality, it can be present at ~1% allele frequency in today's Finnish population, given that the founding bottleneck in Finland occurred just ~100 generations ago. Moreover, out of the 36 Finnish heritage diseases (rare diseases that are relatively common in Finland), 32 of these diseases are autosomal recessive while 2 are autosomal dominant, supporting the hypothesis that deleterious recessive alleles might be enriched in Finland as a result of the bottleneck.

Traditionally, genetic mapping for recessive diseases have been successful in the Finnish population [63,69]. As we demonstrate in Chapter 4, this is both a result of founder alleles being increased in frequency from the bottleneck in Finland, as well as the reduced genetic diversity in rare variation found in founder populations since rare variation is lost as well when undergoing a bottleneck event. Similarly, when we performed genetic mapping in recessive diseases (neuronal ceroid lipofuscinosis and congenital diarrheal disorder) using whole-exome sequencing, we discovered the causal genes (KCTD7 and DGAT1 respectively) using founder populations (Mexicans and Ashkenazi Jewish respectively) relatively easily.

In 2001, a pair of papers first mapped a hexanucleotide repeat expansion in *C9orf72* in the 9p21 region that was associated with risk for amyotrophic lateral sclerosis (ALS) using the Finnish population [70,71]. The authors discovered that the *C9orf72* hexanucleotide repeat expansion accounted for ~46% of familial ALS and ~21% of sporadic ALS within Finland, while the same repeat expansion accounted for only ~6% of familial ALS and ~2% of sporadic ALS in Iranians [72]. This work demonstrated as well that genetic mapping in Finland for monogenic causes in complex diseases can be more fruitful given the reduced background rate of rare

variants in such founder populations. Recently, several papers on the discovery of rare and low-frequency variants associated with complex diseases such as Type 2 diabetes, Alzheimer's disease have also been discovered from other founder populations such as the Icelandic population [73,74,75], highlighting the utility of discovering rare and low-frequency variants in complex diseases from such founder population.

Genetic architecture of the Finns from whole-exome sequencing data

We explored the genetic architecture of the Finns using whole-exome sequencing data in Chapter 4 and showed that Finns are enriched for low-frequency variants with 0.5-5% allele frequency and depleted for extremely rare variants (<0.5% allele frequency). However, across all allele frequency spectrum except for common variants, there are proportionally more loss-of-function variants versus missense and synonymous variants in Finns. In fact, there are almost twice as many low-frequency homozygous loss-of-function variants in an average Finnish individual compared to a non-Finnish European individual. This motivated our large targeted genotyping experiment of 83 low-frequency loss-of-function variants across ~36,000 Finnish individuals in order to discover associations of these low-frequency loss-of-function with a set of quantitative measurements and traits, as well as disease outcomes defined using the medical record system in Finland.

Low-frequency loss-of-function variants in complex diseases and traits

Among the results from our association study, we discovered 2 rare and low-frequency splice variants in the *LPA* gene that lower circulating lipoprotein(a) levels. Previously, increased lipoprotein(a) levels have been associated with increased risk for coronary heart disease [76] and

risk variants in *LPA* have been associated with increased risk for coronary heart disease [77]. However, as in the case of *PCSK9*, it was only until later when compound heterozygous loss-of-function mutations in *PCSK9* were found in a healthy African American individual that suggested knocking out or down the protein levels was suitable as a therapeutic target and did not result in lethality or severe consequences in humans. In our study, we demonstrated that loss-of-function mutations that lower lipoprotein(a) levels are protective against coronary heart disease and that similar to *PCSK9*, we observed several individuals with a complete knockout of the gene but are otherwise healthy. This shows that knocking out or down *LPA* might similar prove to be an effective drug target for coronary heart disease.

SUMMARY

In terms of genetic discoveries for human diseases, this is an incredibly exciting era for geneticists, given the rapid revolutions in technologies, methodologies and approaches for understanding the role of rare variation in diseases. In terms of Mendelian diseases, it is now possible to identify novel genes and variants involved with the use of whole-exome or whole-genome sequencing. And for complex diseases such as autism where genetic discoveries had been extremely difficult, it is now possible to discover specific genes and variants associated with whole-exome sequencing or genotyping.

REFERENCES

1. Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, et al. (1983) A polymorphic DNA marker genetically linked to Huntington's disease. Nature 306: 234-238.

2. HDCRG (1993) A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group. Cell 72: 971-983.

3. Riordan JR, Rommens JM, Kerem B, Alon N, Rozmahel R, et al. (1989) Identification of the cystic fibrosis gene: cloning and characterization of complementary DNA. Science 245: 1066-1073.

4. Morton NE (1955) Sequential tests for the detection of linkage. Am J Hum Genet 7: 277-318.

5. Lander ES, Botstein D (1987) Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. Science 236: 1567-1570.

6. Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265: 2037-2048.

7. Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11: 241-247.

8. Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. Proc Natl Acad Sci U S A 84: 2363-2367.

9. Elston RC, Stewart J (1971) A general model for the genetic analysis of pedigree data. Hum Hered 21: 523-542.

10. Kruglyak L, Daly MJ, Lander ES (1995) Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. Am J Hum Genet 56: 519-527.

11. Yu TW, Chahrour MH, Coulter ME, Jiralerspong S, Okamura-Ikeda K, et al. (2013) Using whole exome sequencing to identify inherited causes of autism. Neuron.

12. Chahrour MH, Yu TW, Lim ET, Ataman B, Coulter ME, et al. (2012) Whole-exome sequencing and homozygosity analysis implicate depolarization-regulated neuronal genes in autism. PLoS Genet 8: e1002635.

13. Morrow EM, Yoo SY, Flavell SW, Kim TK, Lin Y, et al. (2008) Identifying autism loci and genes by tracing recent shared ancestry. Science 321: 218-223.

14. HapMap (2005) A haplotype map of the human genome. Nature 437: 1299-1320.

15. Lee N, Daly MJ, Delmonte T, Lander ES, Xu F, et al. (2001) A genomewide linkage-disequilibrium scan localizes the Saguenay-Lac-Saint-Jean cytochrome oxidase deficiency to 2p16. Am J Hum Genet 68: 397-409.

16. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308: 385-389.

17. Devlin B, Bacanu SA, Roeder K (2004) Genomic Control to the extreme. Nat Genet 36: 1129-1130; author reply 1131.

18. Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467: 832-838.

19. Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52: 506-516.

20. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, et al. (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature 460: 748-752.

21. Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kahler AK, et al. (2013) Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet 45: 1150-1159.

22. Weiss LA, Arking DE, Daly MJ, Chakravarti A (2009) A genome-wide linkage and association scan reveals novel loci for autism. Nature 461: 802-808.

23. Wang K, Zhang H, Ma D, Bucan M, Glessner JT, et al. (2009) Common genetic variants on 5p14.1 associate with autism spectrum disorders. Nature 459: 528-533.

24. Anney R, Klei L, Pinto D, Regan R, Conroy J, et al. (2010) A genome-wide scan for common alleles affecting risk for autism. Hum Mol Genet 19: 4072-4082.

25. Klei L, Sanders SJ, Murtha MT, Hus V, Lowe JK, et al. (2012) Common genetic variants, acting additively, are a major source of risk for autism. Mol Autism 3: 9.

26. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, et al. (2007) Strong association of de novo copy number mutations with autism. Science 316: 445-449.

27. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, et al. (2008) Association between microdeletion and microduplication at 16p11.2 and autism. N Engl J Med 358: 667-675.

28. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, et al. (2011) Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron 70: 863-885.

29. Levy D, Ronemus M, Yamrom B, Lee YH, Leotta A, et al. (2011) Rare de novo and transmitted copy-number variation in autistic spectrum disorders. Neuron 70: 886-897.

30. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461: 272-276.

31. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, et al. (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42: 30-35.

32. Ng SB, Bigham AW, Buckingham KJ, Hannibal MC, McMillin MJ, et al. (2010) Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. Nat Genet 42: 790-793.

33. Haas JT, Winter HS, Lim E, Kirby A, Blumenstiel B, et al. (2012) DGAT1 mutation is linked to a congenital diarrheal disorder. J Clin Invest 122: 4680-4684.

34. Staropoli JF, Karaa A, Lim ET, Kirby A, Elbalalesy N, et al. (2012) A homozygous mutation in KCTD7 links neuronal ceroid lipofuscinosis to the ubiquitin-proteasome system. Am J Hum Genet 91: 202-208.

35. Margolin DH, Kousi M, Chan YM, Lim ET, Schmahmann JD, et al. (2013) Ataxia, dementia, and hypogonadotropism caused by disordered ubiquitination. N Engl J Med 368: 1992-2003.

36. Mannstadt M, Harris M, Bravenboer B, Chitturi S, Dreijerink KM, et al. (2013) Germline mutations affecting Galpha11 in hypoparathyroidism. N Engl J Med 368: 2532-2534.

37. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, et al. (2013) Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med 369: 1502-1511.

38. Abifadel M, Varret M, Rabes JP, Allard D, Ouguerram K, et al. (2003) Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. Nat Genet 34: 154-156.

39. Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, et al. (2005) Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nat Genet 37: 161-165.

40. Kathiresan S (2008) A PCSK9 missense variant associated with a reduced risk of early-onset myocardial infarction. N Engl J Med 358: 2299-2300.

41. Musunuru K, Pirruccello JP, Do R, Peloso GM, Guiducci C, et al. (2010) Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. N Engl J Med 363: 2220-2227.

42. Purcell SM, Moran JL, Fromer M, Ruderfer D, Solovieff N, et al. (2014) A polygenic burden of rare disruptive mutations in schizophrenia. Nature.

43. Chan Y, Lim ET, Sandholm N, Wang SR, McKnight AJ, et al. (2014) An excess of risk-increasing low-frequency variants can be a signal of polygenic inheritance in complex diseases. Am J Hum Genet 94: 437-452.

44. Liu L, Sabo A, Neale BM, Nagaswamy U, Stevens C, et al. (2013) Analysis of rare, exonic variation amongst subjects with autism spectrum disorders and population controls. PLoS Genet 9: e1003443.

45. Heinzen EL, Depondt C, Cavalleri GL, Ruzzo EK, Walley NM, et al. (2012) Exome sequencing followed by large-scale genotyping fails to identify single rare variants of large effect in idiopathic generalized epilepsy. Am J Hum Genet 91: 293-302.

46. O'Roak BJ, Deriziotis P, Lee C, Vives L, Schwartz JJ, et al. (2011) Exome sequencing in sporadic autism spectrum disorders identifies severe de novo mutations. Nat Genet 43: 585-589.

47. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, et al. (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature 485: 242-245.

48. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, et al. (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature.

49. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, et al. (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature.

50. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, et al. (2012) De novo gene disruptions in children on the autistic spectrum. Neuron 74: 285-299.

51. Allen AS, Berkovic SF, Cossette P, Delanty N, Dlugos D, et al. (2013) De novo mutations in epileptic encephalopathies. Nature 501: 217-221.

52. de Ligt J, Willemsen MH, van Bon BW, Kleefstra T, Yntema HG, et al. (2012) Diagnostic exome sequencing in persons with severe intellectual disability. N Engl J Med 367: 1921-1929.

53. Fromer M, Pocklington AJ, Kavanagh DH, Williams HJ, Dwyer S, et al. (2014) De novo mutations in schizophrenia implicate synaptic networks. Nature.

54. Green RC, Berg JS, Grody WW, Kalia SS, Korf BR, et al. (2013) ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. Genet Med 15: 565-574.

55. Lin PI, Kuo PH, Chen CH, Wu JY, Gau SS, et al. (2013) Runs of homozygosity associated with speech delay in autism in a taiwanese han population: evidence for the recessive model. PLoS One 8: e72056.

56. Gamsiz ED, Viscidi EW, Frederick AM, Nagpal S, Sanders SJ, et al. (2013) Intellectual disability is associated with increased runs of homozygosity in simplex autism. Am J Hum Genet 93: 103-109.

57. Novarino G, El-Fishawy P, Kayserili H, Meguid NA, Scott EM, et al. (2012) Mutations in BCKD-kinase lead to a potentially treatable form of autism with epilepsy. Science 338: 394-397.

58. Waterham HR, Koster J, Romeijn GJ, Hennekam RC, Vreken P, et al. (2001) Mutations in the 3beta-hydroxysterol Delta24-reductase gene cause desmosterolosis, an autosomal recessive disorder of cholesterol biosynthesis. Am J Hum Genet 69: 685-694.

59. Svoboda MD, Christie JM, Eroglu Y, Freeman KA, Steiner RD (2012) Treatment of Smith-Lemli-Opitz syndrome and other sterol disorders. Am J Med Genet C Semin Med Genet 160C: 285-294.

60. Tierney E, Conley SK, Goodwin H, Porter FD (2010) Analysis of short-term behavioral effects of dietary cholesterol supplementation in Smith-Lemli-Opitz syndrome. Am J Med Genet A 152A: 91-95.

61. Buchovecky CM, Turley SD, Brown HM, Kyle SM, McDonald JG, et al. (2013) A suppressor screen in Mecp2 mutant mice implicates cholesterol metabolism in Rett syndrome. Nat Genet 45: 1013-1020.

62. Ruzzo EK, Capo-Chichi JM, Ben-Zeev B, Chitayat D, Mao H, et al. (2013) Deficiency of asparagine synthetase causes congenital microcephaly and a progressive form of encephalopathy. Neuron 80: 429-441.

63. Savukoski M, Klockars T, Holmberg V, Santavuori P, Lander ES, et al. (1998) CLN5, a novel gene encoding a putative transmembrane protein mutated in Finnish variant late infantile neuronal ceroid lipofuscinosis. Nat Genet 19: 286-288.

64. Visapaa I, Fellman V, Vesa J, Dasvarma A, Hutton JL, et al. (2002) GRACILE syndrome, a lethal metabolic disorder with iron overload, is caused by a point mutation in BCS1L. Am J Hum Genet 71: 863-876.

65. Pajukanta P, Lilja HE, Sinsheimer JS, Cantor RM, Lusis AJ, et al. (2004) Familial combined hyperlipidemia is associated with upstream transcription factor 1 (USF1). Nat Genet 36: 371-376.

66. Montague CT, Farooqi IS, Whitehead JP, Soos MA, Rau H, et al. (1997) Congenital leptin deficiency is associated with severe early-onset obesity in humans. Nature 387: 903-908.

67. Barak T, Kwan KY, Louvi A, Demirbilek V, Saygi S, et al. (2011) Recessive LAMC3 mutations cause malformations of occipital cortical development. Nat Genet 43: 590-594.

68. Bitner-Glindzicz M, Lindley KJ, Rutland P, Blaydon D, Smith VV, et al. (2000) A recessive contiguous gene deletion causing infantile hyperinsulinism, enteropathy and deafness identifies the Usher type 1C gene. Nat Genet 26: 56-60.

69. Vesa J, Hellsten E, Verkruyse LA, Camp LA, Rapola J, et al. (1995) Mutations in the palmitoyl protein thioesterase gene causing infantile neuronal ceroid lipofuscinosis. Nature 376: 584-587.

70. Renton AE, Majounie E, Waite A, Simon-Sanchez J, Rollinson S, et al. (2011) A hexanucleotide repeat expansion in C9ORF72 is the cause of chromosome 9p21-linked ALS-FTD. Neuron 72: 257-268.

71. DeJesus-Hernandez M, Mackenzie IR, Boeve BF, Boxer AL, Baker M, et al. (2011) Expanded GGGGCC hexanucleotide repeat in noncoding region of C9ORF72 causes chromosome 9p-linked FTD and ALS. Neuron 72: 245-256.

72. Alavi A, Nafissi S, Rohani M, Shahidi G, Zamani B, et al. (2014) Repeat expansion in C9ORF72 is not a major cause of amyotrophic lateral sclerosis among Iranian patients. Neurobiol Aging 35: 267 e261-267.

73. Jonsson T, Atwal JK, Steinberg S, Snaedal J, Jonsson PV, et al. (2012) A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. Nature 488: 96-99.

74. Steinthorsdottir V, Thorleifsson G, Sulem P, Helgason H, Grarup N, et al. (2014) Identification of low-frequency and rare sequence variants associated with elevated or reduced risk of type 2 diabetes. Nat Genet.

75. Jonsson T, Stefansson H, Steinberg S, Jonsdottir I, Jonsson PV, et al. (2013) Variant of TREM2 associated with the risk of Alzheimer's disease. N Engl J Med 368: 107-116.

76. Erqou S, Kaptoge S, Perry PL, Di Angelantonio E, Thompson A, et al. (2009) Lipoprotein(a) concentration and the risk of coronary heart disease, stroke, and nonvascular mortality. JAMA 302: 412-423.

77. Clarke R, Peden JF, Hopewell JC, Kyriakou T, Goel A, et al. (2009) Genetic variants associated with Lp(a) lipoprotein level and coronary disease. N Engl J Med 361: 2518-2528.

CHAPTER 2

Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders

Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders

Elaine T. Lim[1,4,5,6,7], Soumya Raychaudhuri[4,6,9], Stephan J. Sanders[10], Christine Stevens[4], Aniko Sabo[11], Daniel G. MacArthur[1,4,6], Benjamin M. Neale[1,4,5,6], Andrew Kirby[1,4,6], Douglas M. Ruderfer[1,3,4,5,6,8,12], Menachem Fromer[1,3,4,5,6,8,12], Monkol Lek[1,4,6], Li Liu[18], Jason Flannick[1,2,4,6], Stephan Ripke[1,4,5], Uma Nagaswamy[11], Donna Muzny[11], Jeffrey G. Reid[11], Alicia Hawes[11], Irene Newsham[11], Yuanqing Wu[11], Lora Lewis[11], Huyen Dinh[11], Shannon Gross[11], Li-San Wang[19], Chiao-Feng Lin[19], Otto Valladares[19], Stacey B. Gabriel[4], Mark dePristo[4], David M. Altshuler[1,2,4,6], Shaun M. Purcell[1,3,4,5,6,8,12], NHLBI Exome Sequencing Project, Matthew W. State[10], Eric Boerwinkle[11,21], Joseph D. Buxbaum[13,14,15,16,17], Edwin H. Cook[22], Richard A. Gibbs[11], Gerard D. Schellenberg[20], James S. Sutcliffe[23], Bernie Devlin[24], Kathryn Roeder[18], and Mark J. Daly[1,4,5,6,*]

[1] Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA.

[2] Department of Molecular Biology, Massachusetts General Hospital, Boston, MA 02114, USA.

[3] Psychiatric & Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA.

[4] Program in Medical and Population Genetics, Broad Institute, Cambridge, MA 02142, USA.

[5] Stanley Center for Psychiatric Research, Broad Institute, Cambridge, MA 02142, USA.

[6] Departments of Genetics and Medicine, Harvard Medical School, Boston, MA 02115, USA.

[7] Program in Genetics and Genomics, Biological and Biomedical Sciences, Harvard Medical School, Boston, MA 02115, USA.

[8] Department of Psychiatry, Harvard Medical School, Boston, MA 02115, USA.

[9] Division of Immunology, Allergy, and Rheumatology, Brigham and Women's Hospital, Boston, MA 02115, USA.

[10] Departments of Psychiatry and Genetics, Yale University School of Medicine, New Haven, CT 06520, USA.

[11] Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA.

[12] Division of Psychiatric Genomics, Mount Sinai School of Medicine, New York, NY 10029, USA.

[13] Seaver Autism Center for Research and Treatment, Mount Sinai School of Medicine, New York, NY 10029, USA.

[14] Department of Psychiatry, Mount Sinai School of Medicine, New York, NY 10029, USA.

[15] Department of Genetics and Genomic Sciences, Mount Sinai School of Medicine, New York, NY 10029, USA.

[16] Department of Neuroscience, Mount Sinai School of Medicine, New York, NY 10029, USA.

[17] Friedman Brain Institute, Mount Sinai School of Medicine, New York, NY 10029, USA.

[18] Department of Statistics and Lane Center for Computational Biology, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

[19] Penn Center for Bioinformatics, University of Pennsylvania, Philadelphia, PA 19104, USA.

[20] Pathology and Laboratory Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA.

[21] Human Genetics Center, University of Texas Health Science Center at Houston, TX 77030, USA.

[22] Department of Psychiatry, University of Illinois at Chicago, Chicago, IL 60612, USA.

[23] Departments of Molecular Physiology & Biophysics and Psychiatry, Vanderbilt Brain Institute, Vanderbilt University, Nashville, TN 37232, USA.

[24] Department of Psychiatry, University of Pittsburgh School of Medicine, Pittsburgh, PA 15260, USA.

[*] Correspondence: mjdaly@atgu.mgh.harvard.edu

Originally published as:

Lim ET *et al*. Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. Neuron, 2013. 77(2): p.235-42.

ABSTRACT

To characterize the role of rare complete human knockouts in autism spectrum disorders (ASD), we identify genes with homozygous or compound heterozygous loss-of-function (LoF) variants (defined as nonsense and essential splice sites) from exome sequencing of 933 cases and 869 controls. We identify a two-fold increase in complete knockouts of autosomal genes with low rates of LoF variation (≤5% frequency) in cases and estimate a 3% contribution to ASD risk by these events, confirming this observation in an independent set of 563 probands and 4,605 controls. Outside the pseudo-autosomal regions on the X-chromosome, we similarly observe a significant 1.5-fold increase in rare hemizygous knockouts in males, contributing to another 2% of ASDs in males. Taken together these results provide compelling evidence that rare autosomal and X-chromosome complete gene knockouts are important inherited risk factors for ASD.

INTRODUCTION

Autism spectrum disorder (ASD) is a highly heritable, common disorder that affects ~1 in 88 individuals [1]. Previous studies have shown a reproducible contribution of *de novo* copy number variants (CNVs) [2,3,4,5,6] and *de novo* single nucleotide variants (SNVs) [7,8,9,10] to ASD risk - though these effects provide little explanation for the widely recognized high heritability [11]. An early segregation analysis on 46 multiplex families (each with multiple affected children) suggested evidence for an autosomal recessive (or '2-hit') model in ASD [12] with a subsequent study showing that ASD is unlikely to fit a model with a major gene effect [13]. Further to this point, the most recent results from *de novo* CNVs and SNVs point to a model in which hundreds of genes are likely to contribute to autism risk. Building from these observations, as a means of providing insight into the heritable component of ASD risk, we

sought to test the hypothesis that 2-hit etiologies exist in ASD and that these events, like the *de novo* CNVs and SNVs, are most likely to be distributed over many genes. Supporting this hypothesis are historical segregation analyses [12,14], the successful use of homozygosity mapping in consanguineous populations [15], as well as recent studies showing that ASD probands had a significant excess of homozygous haplotype sharing, suggesting that there are recessive loci in these risk-conferring haplotypes [16,17]. Other studies have also implicated the role of a 2-hit or oligogenic model for rare CNVs in ASD [18].

It has been shown that there are relatively few homozygous or compound heterozygous LoF variants (i.e., complete gene knockouts) in healthy individuals. Most of these complete knockouts found are common (MAF>5%) and are distributed across a very small number (~100-200) of genes, such as the olfactory receptors, that are apparently inessential and do not result in any obvious phenotype or severe medical consequence [19]. We similarly observe in these ASD datasets that an average individual harbors ~5 common complete knockouts (from nonsense and essential splice site variants) distributed across a small subset of genes on the autosomes. In striking contrast, if we consider only LoF variants with frequency ≤5%, fewer than 5% of individuals harbor even a single rare complete knockout (Table 2.1). While heterozygous LoF mutations are seen in thousands of genes, the very low frequency and paucity of observed complete knockouts suggests a broad pool of genes (including many Mendelian disorders) where 2-hit variants may give rise to severe and reproductively deleterious phenotypes. While genes with common complete knockouts are more likely to be benign (or unlikely to result in severe phenotypes with high penetrance), genes with rare complete knockouts are more likely to be disease-causing [20] simply because selection prevents deleterious recessive-acting variants from reaching even moderate allele frequencies.

Table 2.1: Population Distribution of Rare and Common LoFs

The average number of rare (≤5%) and common (>5%) homozygous LoF variants, as well as the average number of such variants calculated from the BI case-control dataset.

	Average number of homozygous variants	Number of unique genes with a homozygous variant	Average number of heterozygous variants	Number of unique genes with a heterozygous variant
Rare (≤5%) LoFs	0.05 variant per individual	33 genes	13 variants per individual	3,409 genes
Common (>5%) LoFs	5 variants per individual	96 genes	36 variants per individual	99 genes

---

If a subset of ASD cases were caused by rare 2-hit events with large effects (e.g. odds ratios of >5) distributed across many different genes, then family-based linkage or GWAS would have little power to detect such events, as each locus individually would explain a very small fraction of all cases given the commonness of the outcome and the large number of ASD genes. To evaluate evidence for such 2-hit etiologies in ASD, we studied the distribution and patterns of rare complete knockouts from whole-exome sequence data across two case-control studies comprised of 1,802 European subjects to identify events in which individuals carried 2 LoF autosomal variants in a single gene *in trans*. In this study, we show that rare complete knockouts

on the autosomes (variant allele frequencies of ≤5%) are significantly enriched in cases, suggesting that these events contribute to the genetic etiology of ASD.

A variant with a diploid allele frequency of 5% on the autosomes results in a complete knockout in 0.25% of the individuals. Outside the pseudo-autosomal regions on the X-chromosome in males, a single LoF variant with 0.25% allele frequency also results in a complete knockout in 0.25% of males. Similarly, we found that rare complete knockouts on the X-chromosome (variant allele frequencies of ≤0.25%) are also significantly enriched in male cases, further reinforcing the role of rare complete knockouts as risk factors for ASD.

RESULTS

Exome Capture and Sequencing

To assess the contribution of rare complete knockouts to ASD, we analyzed data from an ethnically-matched case-control population. We selected 933 cases and 869 controls sequenced in this study by matching them with multi-dimensional scaling (MDS) of common variants genotyped on Illumina 1M, Affymetrix 5.0, or 6.0 arrays [21] to reduce potential confounding by population stratification. The exomes were sequenced at two different sequencing centers – the Broad Institute (BI) and the Baylor College of Medicine (BCM). A total of 428 ASD cases selected from the Autism Genetic Resource Exchange (AGRE) and 378 NIMH controls (a total of 806 individuals) were sequenced at BI, and another 505 ASD cases selected from the Autism Simplex Collection (TASC) and 491 NIMH controls (a total of 996 individuals) were sequenced at BCM, resulting in 1,802 individuals across the two case-control datasets. All controls were selected from an NIMH control repository and were ascertained for not having schizophrenia or bipolar mood disorder. Another 563 probands were added into the final analyses (388

trios/quartets from the Simons Simplex Collection (SSC) [8,10], 175 trios from the Boston Autism Consortium sequenced at BI (104 from [9]) and together with 4,605 additional European controls from the NHLBI exome sequencing project and the 1000 Genomes Project, this resulted in a total of >6,000 exomes used in this study (Table 2.2). The metrics for the case-control datasets are described in Table 2.3.

Enrichment of Rare Complete Knockouts in ASD

Given that rare complete knockouts consist of both compound heterozygous and homozygous variants on the autosomes, we adapted a statistical phasing approach similar to the four-haplotype test to eliminate instances in which multiple LoF variants may segregate *in cis* (Figure 2.1). There are a total of 91 such rare complete knockouts in the case-control datasets, with 62 of these found in the cases compared to 29 in the controls (Table 2.4), representing a roughly 2-fold enrichment of these events in the cases (odds ratio (OR) = 2.0, 95% CI = [1.5, 2.5], one-sided permutation $P = 0.0017$). Based on the difference between cases and controls (6% of the cases versus 3.3% of the controls have a rare complete knockout), we estimate a ~3% contribution by rare complete knockouts to ASD. While different capture and sequencing technologies were employed at the two sequencing centers, and different depths of sequencing achieved (Liu et al., personal communication), the excess in cases was consistent in the two datasets (ORs = 2.1, 95% CI = [1.5, 2.7] and 1.8, 95% CI = [1.1, 2.5]).

Table 2.2: Description of Available Datasets

The case-control datasets comprise of cases and controls that were well-matched for their common variants and sequenced at the BI and BCM were used to assess the enrichment of rare complete knockouts (Datasets 1 and 2). For the final analyses, probands from 200 quartets sequenced at Yale University, 188 quartets sequenced at Cold Spring Harbor Laboratory (CSHL), 104 published trios and 71 new trios from BI were compared to another 4,605 controls from the NHLBI Exome Sequencing Project and the 1000 Genomes Project (Datasets 3-8), resulting in a total of 1,496 cases and 5,474 controls in this study.

	Type	Sequencing Center	Number of cases	Number of controls
Dataset 1	Case-control	BI (unpublished)	428 cases	378 controls
Dataset 2	Case-control	BCM (unpublished)	505 cases	491 controls
Total for case-control datasets			**933 cases**	**869 controls**
Dataset 3	Quartets	Yale [10]	200 probands	-
Dataset 4	Quartets	CSHL [8]	188 probands	-
Dataset 5	Trios	BI [9]	104 probands	-
Dataset 6	Trios	BI (unpublished)	71 probands	-

Table 2.2: Description of Available Datasets (Continued)

	Type	Sequencing Center	Number of cases	Number of controls
Dataset 7	Controls	Multiple centers	-	4419 controls (NHLBI Exome Sequencing Project)
Dataset 8	Controls	Multiple centers	-	186 controls (1000 Genomes Project)
Total numbers for final analyses			**1496 cases**	**5474 controls**

Table 2.3: Metrics for BI and BCM Case-Control Datasets

(A) The number of genes, variants and average coverage in the BI and BCM case-control datasets.

(B) Number of variants and the distribution of the variants in both the BI and BCM case-control datasets.

A

	# Genes	Average minor allele frequency	# Variants per gene	Average depth
BCM	15759	0.021	11.28	62.41
BI	17900	0.024	18.49	140.22

B

	# Variants	# Missense variants	# Nonsense variants	# Synonymous variants	# Singletons	Common variants (>5% MAF)	Rare variants (≤5% MAF)
BCM	177699	95522	2190	61720	111766	18231	159468
BI	330985	198070	4149	128766	171813	40080	290897

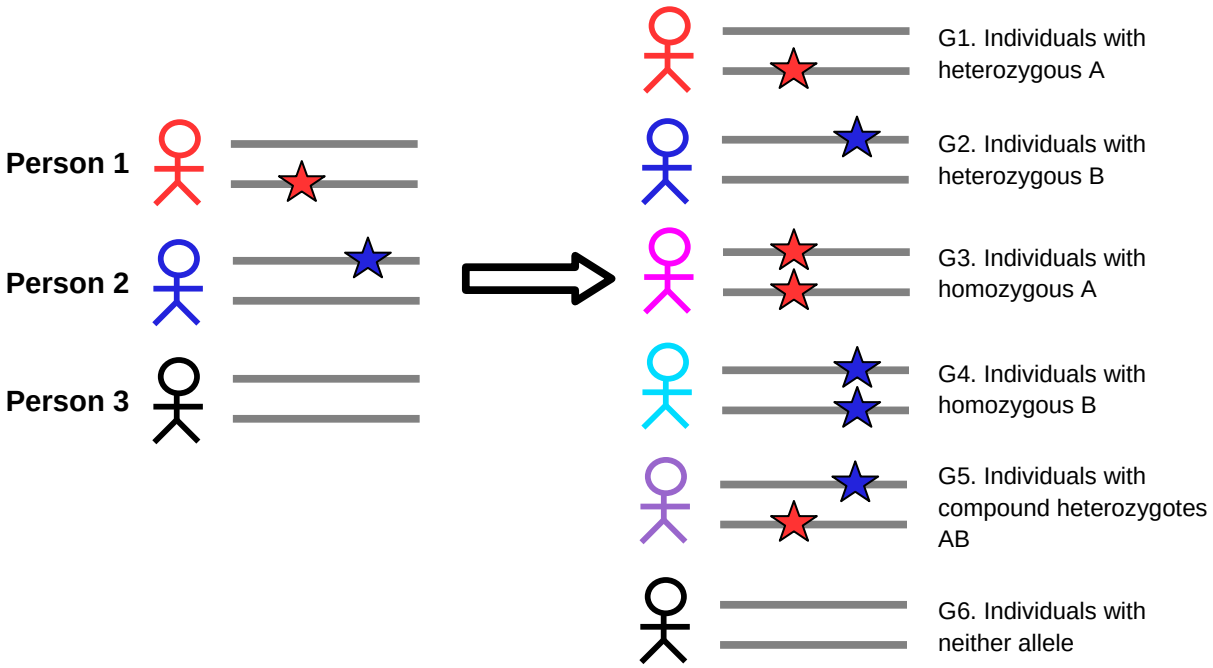
Figure 2.1: LD-based Phasing Approach To Predict Compound Heterozygous Variants

(A) If both the red and blue variants within the same gene are *in trans* (they occurred on different chromosomes in Persons 1 and 2), we will observe some individuals in the population with only the red variant (G1) and some individuals with only the blue variant (G2).

(B) If the blue variant occurred on the same chromosome as the red variant within the same gene at some point in time and is *in cis* with the red variant, we will observe some individuals in the population with only the red variant (G7), but will not observe any individual with only the blue variant (G2). In addition, there might be some individuals with 3 or 4 copies of alleles across both variants (G10 and G11), suggesting that both variants are more likely to be *in cis*.

Figure 2.1: LD-based Phasing Approach To Predict Compound Heterozygous Variants (Continued)

A



B

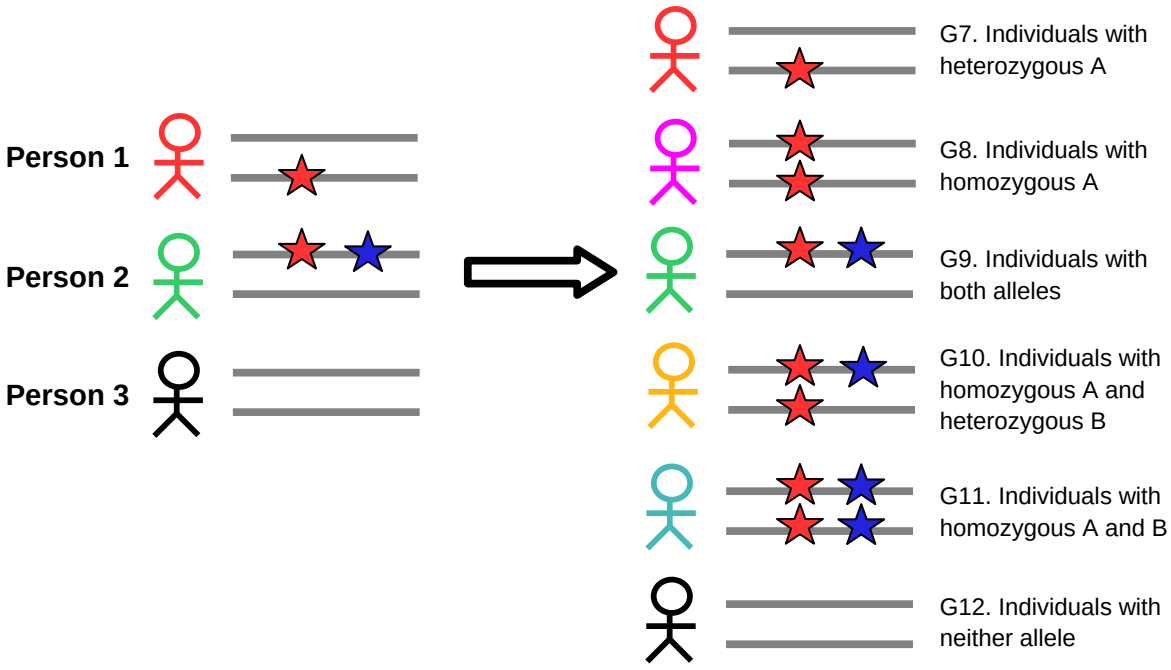


Table 2.4: Number of 2-hit Events Observed in the Cases and Controls

(A) Number of 2-hit events observed in the cases compared to controls for the various allele frequencies and various functional categories: LoF, non-synonymous (NS) and synonymous (SYN) and the numbers in brackets indicate the number of homozygous events. The difference between the total number of 2-hit events and the homozygous events are the compound heterozygous events.

(B) Number of heterozygous variants observed in the cases compared to controls for the various allele frequencies and various functional categories: LoF, non-synonymous (NS) and synonymous (SYN).

A

Allele Freq	# 2-hit LoF in cases	# 2-hit LoF in controls	# 2-hit NS in cases	# 2-hit NS in controls	# 2-hit SYN in cases	# 2-hit SYN in controls
≤5%	62 (42)	29 (19)	18857 (3465)	15365 (2783)	13417 (3513)	10034 (2548)
≤10%	172 (142)	149 (121)	38115 (12345)	32199 (10283)	31743 (13261)	25943 (10900)
≤20%	617 (585)	538 (505)	90237 (43948)	78646 (38339)	88421 (51844)	76202 (45036)
≤30%	1024 (988)	876 (840)	158175 (93604)	139520 (82719)	176497 (119533)	155521 (106061)
≤40%	1645 (1603)	1418 (1378)	242024 (162019)	215878 (144892)	287349 (214735)	255106 (191552)

Table 2.4: Number of 2-hit Events Observed in the Cases and Controls (Continued)

A

Allele Freq	# 2-hit LoF in cases	# 2-hit LoF in controls	# 2-hit NS in cases	# 2-hit NS in controls	# 2-hit SYN in cases	# 2-hit SYN in controls
≤50%	3331 (3240)	3029 (2953)	315502 (245761)	280425 (220735)	395881 (317715)	351109 (283265)

B

Allele Freq	# LoF variants in cases	# LoF variants in controls	# NS variants in cases	# NS variants in controls	# SYN variants in cases	# SYN variants in controls
≤5%	9365	8400	517337	451650	428510	367373
≤10%	12479	11230	817404	728463	747665	660771
≤20%	19472	17664	1396259	1262890	1437932	1294969
≤30%	24355	20386	1981964	1803775	2218066	2015026
≤40%	29198	26749	2567749	2342964	3409983	2748908
≤50%	35087	32289	3133619	2863937	3821090	3486579

Using the results from a previous study of expression patterns of post-mortem brains [22], we observed the enrichment in rare complete knockouts in cases was particularly pronounced in genes found to be expressed in the brain, with 37 events in cases compared to only 13 in the controls (OR = 2.7, 95% CI = [2.1, 3.3], one-sided permutation $P = 0.002$), although this enrichment in brain-expressed genes was not significantly different from the global enrichment observed (one-sided permutation $P = 0.13$, Figure 2.2).

To confirm that this excess was not an artifact of any residual uncertainty in statistical phasing, we examined the subset of rare complete knockouts that were homozygous LoF variants alone and found that these events were also significantly enriched by 2-fold (42 in cases and 19 in controls, OR = 2.1, 95% CI = [1.6, 2.6], one-sided permutation $P = 0.0059$, Table 2.4). We further ensured that the excess was not driven by inaccuracies in phasing 'singleton' variants (variants that were observed only once in a single individual) and found that rare complete knockouts excluding the singleton variants were also significantly enriched (48 in cases and 24 in controls, OR = 1.9, 95% CI = [1.4, 2.4], one-sided permutation $P = 0.0081$). Since an excess in 2-hit LoF could arise trivially if there was a significant overall difference in rates of LoF variants between cases and controls, we evaluated the total number of single-copy losses (i.e., heterozygous LoF carriers) with variant allele frequencies ≤5% found in cases compared to controls and saw no enrichment (OR = 1.0, 95% CI = [0.9, 1.1], Table 2.5).

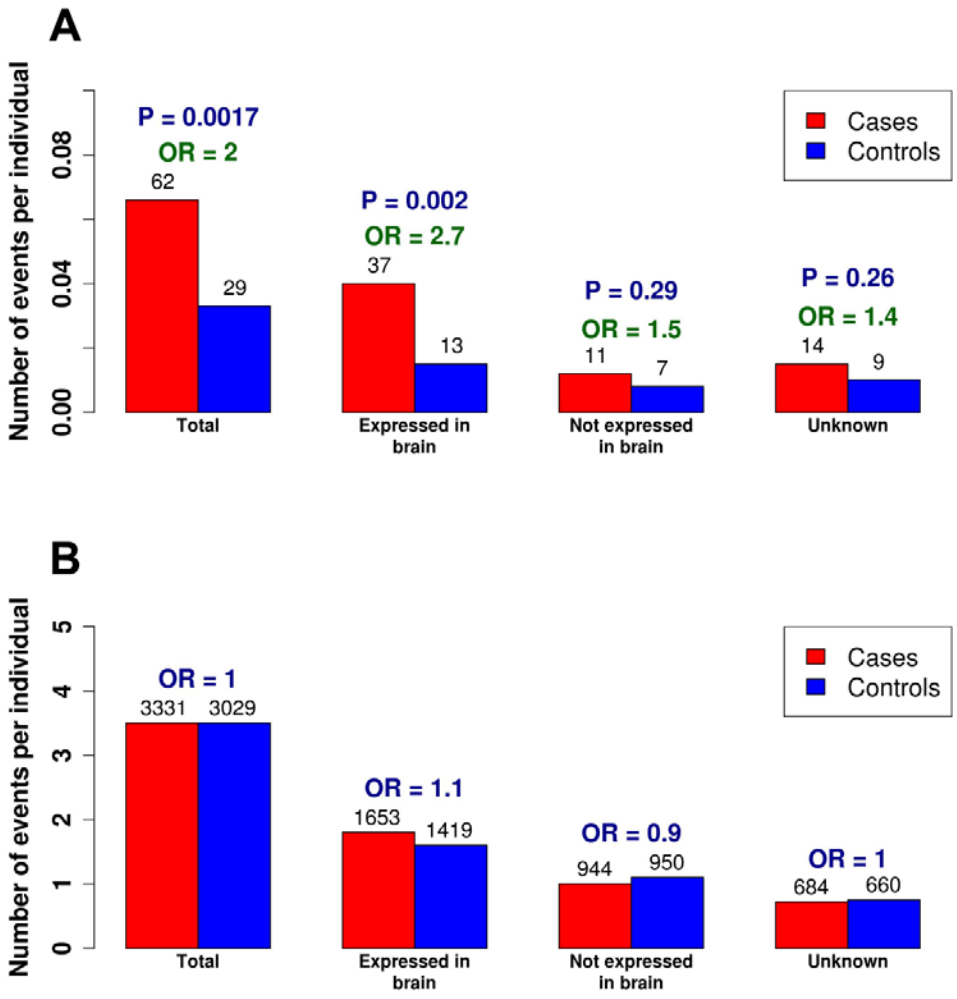


Figure 2.2: Expression Patterns of the Complete Knockouts

(A) The enrichment of rare complete knockouts in cases versus controls.

(B) The enrichment observed in rare complete knockouts is not observed in the common complete knockouts.

The x-axis indicates the average number of events per individual in cases and controls and the numbers above the barplots indicate the total number of such events in cases and controls, with the odds ratios (OR) shown above.

Table 2.5: Control Experiments for the Enrichment Observed in Rare Complete Knockouts

(A) The ratio of rare complete knockouts found in cases versus controls is significantly different from the ratio of rare 2-hit synonymous variants in cases versus controls.

(B) The ratio of rare complete knockouts found in cases versus controls is also significantly different from the ratio of common complete knockouts in cases versus controls.

The hypergeometric probabilities reflect the probabilities of the enrichment observed in rare complete knockouts after correcting for the enrichment in rare 2-hit synonymous and common complete knockouts.

A

	Not expressed in brain	Expressed in brain	Unknown	Total
Cases	1,172	10,894	1,351	13,417
Controls	873	8,196	965	10,034
Odds ratio of rare 2-hit synonymous events	1.3	1.2	1.3	1.3
Odds ratio of rare 2-hit LoF events	1.5	2.7	1.4	2
Hypergeometric *P*	0.47	0.01	0.49	0.021

Table 2.5: Control Experiments for the Enrichment Observed in Rare Complete Knockouts (Continued)

B

	Not expressed in brain	Expressed in brain	Unknown	Total
Cases	994	1,653	684	3,331
Controls	950	1,419	660	3,029
Odds ratio of common 2-hit LoF events	1	1.1	1	1
Odds ratio of rare 2-hit LoF events	1.5	2.7	1.4	2
Hypergeometric *P*	0.27	0.0025	0.24	0.0015

Finally, we validated all variants by ensuring that they were either present in dbSNP, the NHLBI Exome Sequencing Project and/or were confirmed using Fluidigm genotyping, Sanger sequencing or Fluidigm PCR with MiSeq sequencing with 94% of these variants validating as true polymorphisms (Table 2.6**,** Table 2.7). Even conservatively assuming all validation failures were false positive SNPs (rather than genotyping assay failures), removing the three events in cases and two in controls from the overall tallies has no impact on the results. As a final check, we used rare homozygous and compound heterozygous (or '2-hit') synonymous events, as well as common complete knockouts, as internal controls and confirmed the enrichment of rare complete knockouts was far greater and significantly different compared to both of these (Table 2.5).

Knockouts via homozygosity of rare LoF sites could arise from hemizygous LoF variants that were exposed through the deletion of the other copy in the gene region. Using a CNV-calling algorithm for exome sequencing (XHMM) [23], we found that 2 of the homozygous LoFs observed in cases (E201X in KRT83 and E211X in PRAMEF2) were, in fact, LoF variants unmasked by deletions spanning across the regions (11kb and 183kb deletions respectively), although this does not change the fact that they are complete gene knockouts.

To confirm these observations, we examined an independent set of cases (N = 563) from recent trio sequencing efforts (where 2-hit knockout status was certain from the existence of parental sequence data) and compared to a broader population dataset (N = 4,605) from the NHLBI exome sequencing project and 1000 Genomes Project (Table 2.2). The enrichment (7.6% in cases to 5.5% in controls, hypergeometric test P = 0.016) was replicated in this comparison as well – further confirming the veracity of this observation.

Table 2.6: List of Autosomal Genes with Rare Complete Knockouts

Gene	Variant 1 Annotation	Variant 2 Annotation	# Cases	# Controls	Variant 1 found in dbSNP?	Variant 1 found in NHLBI ESP?	Variant 2 found in dbSNP?	Variant 2 found in NHLBI ESP?	# Individuals validated for Variant 1	# Individuals validated for Variant 2	*In trans?*
C1orf127	G->A	G->A	1	0	Y	Y	Y	Y	1	-	-
PRAMEF2	E211*	E211*	1	0	Y	Y	Y	Y	1	-	-
PGM1	G->A	G->A	0	1	Y	Y	Y	Y	0	-	-
FAM71A	K555*	K555*	1	0	Y	Y	Y	Y	0	-	-
C2orf53	S78*	S78*	0	1	N	N	N	N	Did not validate	-	-
C2orf63	A->T	A->T	0	1	Y	Y	Y	Y	0	-	-
VWA3B	E372*	E372*	1	0	Y	Y	Y	Y	1	-	-
DPP4	T->C	T->C	2	1	Y	Y	Y	Y	2	-	-
IFIH1	C->G	C->G	1	0	Y	Y	Y	Y	1	-	-
PTH2R	S82*	S82*	1	0	Y	N	Y	N	1	-	-
ATP13A5	Q355*	Q355*	0	1	Y	Y	Y	Y	0	-	-
CC2D2A	R88*	R88*	0	1	Y	Y	Y	Y	0	-	-

Table 2.6: List of Autosomal Genes with Rare Complete Knockouts (Continued)

Gene	Variant 1 Annotation	Variant 2 Annotation	# Cases	# Controls	Variant 1 found in dbSNP?	Variant 1 found in NHLBI ESP?	Variant 2 found in dbSNP?	Variant 2 found in NHLBI ESP?	# Individuals validated for Variant 1	# Individuals validated for Variant 2	*In trans?*
UGT2A1	Y192*	Y192*	1	0	Y	Y	Y	Y	1	-	-
C1orf127	G->A	G->A	1	0	Y	Y	Y	Y	1	-	-
PRAMEF2	E211*	E211*	1	0	Y	Y	Y	Y	1	-	-
PGM1	G->A	G->A	0	1	Y	Y	Y	Y	0	-	-
FAM71A	K555*	K555*	1	0	Y	Y	Y	Y	0	-	-
C2orf53	S78*	S78*	0	1	N	N	N	N	Did not validate	-	-
C2orf63	A->T	A->T	0	1	Y	Y	Y	Y	0	-	-
VWA3B	E372*	E372*	1	0	Y	Y	Y	Y	1	-	-
DPP4	T->C	T->C	2	1	Y	Y	Y	Y	2	-	-
IFIH1	C->G	C->G	1	0	Y	Y	Y	Y	1	-	-
PTH2R	S82*	S82*	1	0	Y	N	Y	N	1	-	-
ATP13A5	Q355*	Q355*	0	1	Y	Y	Y	Y	0	-	-
CC2D2A	R88*	R88*	0	1	Y	Y	Y	Y	0	-	-

Table 2.6: List of Autosomal Genes with Rare Complete Knockouts (Continued)

Gene	Variant 1 Annotation	Variant 2 Annotation	# Cases	# Controls	Variant 1 found in dbSNP?	Variant 1 found in NHLBI ESP?	Variant 2 found in dbSNP?	Variant 2 found in NHLBI ESP?	# Individuals validated for Variant 1	# Individuals validated for Variant 2	*In trans?*
UGT2A1	Y192*	Y192*	1	0	Y	Y	Y	Y	1	-	-
AGXT2	C->T	C->T	1	0	Y	Y	Y	Y	0	-	-
FAM81B	Q144*	Q144*	1	1	Y	Y	Y	Y	0	-	-
SLC17A4	Q433*	Q433*	0	1	Y	Y	Y	Y	0	-	-
PNPLA1	Y488*	Y488*	1	1	Y	Y	Y	Y	0	-	-
USP45	A->G	A->G	0	1	Y	Y	Y	Y	0	-	-
KIAA1919	L182*	L182*	1	0	Y	N	Y	N	1	-	-
TAAR2	W168*	W168*	1	0	Y	Y	Y	Y	1	-	-
LPA	T->C	T->C	1	0	Y	Y	Y	Y	1	-	-
UNC93A	G->C	G->C	1	0	Y	Y	Y	Y	1	-	-
RNF32	T->C	T->C	2	0	Y	Y	Y	Y	2	-	-
LRRC69	T->A	T->A	1	0	Y	Y	Y	Y	0	-	-
PKHD1L1	G->A	G->A	0	1	Y	Y	Y	Y	0	-	-
ZNF883	R341*	R341*	0	1	Y	Y	Y	Y	0	-	-

Table 2.6: List of Autosomal Genes with Rare Complete Knockouts (Continued)

Gene	Variant 1 Annotation	Variant 2 Annotation	# Cases	# Controls	Variant 1 found in dbSNP?	Variant 1 found in NHLBI ESP?	Variant 2 found in dbSNP?	Variant 2 found in NHLBI ESP?	# Individuals validated for Variant 1	# Individuals validated for Variant 2	*In trans?*
OR10V1	Q123*	Q123*	1	0	Y	Y	Y	Y	1	-	-
RAD52	Y415*	Y415*	1	0	Y	Y	Y	Y	1	-	-
KRT83	E201*	E201*	1	1	Y	Y	Y	Y	1	-	-
OAS3	R812*	R812*	1	0	N	N	N	N	1	-	-
OLFM4	R214*	R214*	0	1	Y	Y	Y	Y	0	-	-
CLYBL	R259*	R259*	2	0	Y	Y	Y	Y	0	-	-
C14orf105	Q183*	Q183*	1	0	Y	Y	Y	Y	1	-	-
NPC2	C->T	C->T	0	1	Y	Y	Y	Y	0	-	-
SERPINA10	R88*	R88*	1	0	Y	Y	Y	Y	0	-	-
RAGE	R94*	R94*	0	1	Y	Y	Y	Y	0	-	-
ABCC12	W1024*	W1024*	2	1	Y	Y	Y	Y	1	-	-
DBF4B	G->T	G->T	1	0	Y	Y	Y	Y	0	-	-
C17orf57	R211*	R211*	1	0	Y	Y	Y	Y	0	-	-
ABCA10	R1322*	R1322*	4	0	Y	Y	Y	Y	4	-	-

Table 2.6: List of Autosomal Genes with Rare Complete Knockouts (Continued)

Gene	Variant 1 Annotation	Variant 2 Annotation	# Cases	# Controls	Variant 1 found in dbSNP?	Variant 1 found in NHLBI ESP?	Variant 2 found in dbSNP?	Variant 2 found in NHLBI ESP?	# Individuals validated for Variant 1	# Individuals validated for Variant 2	*In trans?*
FLJ35220	A->G	A->G	1	1	Y	Y	Y	Y	1	-	-
STARD6	R19*	R19*	5	0	Y	Y	Y	Y	3	-	-
ZNF77	Q100*	Q100*	1	0	Y	Y	Y	Y	1	-	-
UNC13A	G->A	G->A	1	0	Y	Y	Y	Y	1	-	-
ZNF780B	T->C	T->C	1	0	Y	Y	Y	Y	1	-	-
LAIR2	R76*	R76*	0	1	Y	Y	Y	Y	0	-	-
ZIM3	K438*	K438*	0	1	Y	Y	Y	Y	0	-	-
KIAA1755	R510*	R510*	0	1	Y	Y	Y	Y	0	-	-
C1orf168	T->C	K198*	1	0	Y	Y	N	N	1	1	N
DNAH14	L286*	E3391*	0	1	N	N	N	Y	1	1	-
VWA3B	E372*	A->T	1	0	Y	Y	N	N	1	Did not validate	-
DNAH7	Y3978*	R1009*	1	0	Y	Y	N	Y	1	1	-
GBE1	A->G	Q46*	0	1	Y	Y	N	N	1	1	-

Table 2.6: List of Autosomal Genes with Rare Complete Knockouts (Continued)

Gene	Variant 1 Annotation	Variant 2 Annotation	# Cases	# Controls	Variant 1 found in dbSNP?	Variant 1 found in NHLBI ESP?	Variant 2 found in dbSNP?	Variant 2 found in NHLBI ESP?	# Individuals validated for Variant 1	# Individuals validated for Variant 2	*In trans?*
ATP13A5	A->G	Q355*	1	0	N	N	Y	Y	1	1	-
CC2D2A	R88*	A->T	0	1	Y	Y	N	N	1	Did not validate	-
WDR17	A->T	A->T	1	0	N	N	N	N	Did not validate	Did not validate	-
INTS8	A->T	A->T	1	0	N	Y	N	N	Did not validate	Did not validate	-
SLC22A25	T->C	W120*	1	0	Y	Y	N	N	1	1	Y
MMP1	C->T	A->G	0	1	Y	Y	Y	Y	0	0	-
RAD52	Y415*	S346*	1	0	Y	Y	Y	Y	1	1	-
C12orf64	G->T	R1808*	1	0	N	N	N	N	1	1	-
ZSCAN29	Q669*	C645*	1	0	N	N	N	N	1	1	-
TMC3	S1045*	R736*	1	0	Y	Y	N	Y	0	0	-
C17orf57	R211*	G->A	1	0	Y	Y	Y	Y	1	1	Y

Table 2.6: List of Autosomal Genes with Rare Complete Knockouts (Continued)

Gene	Variant 1 Annotation	Variant 2 Annotation	# Cases	# Controls	Variant 1 found in dbSNP?	Variant 1 found in NHLBI ESP?	Variant 2 found in dbSNP?	Variant 2 found in NHLBI ESP?	# Individuals validated for Variant 1	# Individuals validated for Variant 2	*In trans?*
C17orf57	R211*	K433*	0	1	Y	Y	Y	Y	0	0	-
C17orf57	R211*	Y546*	2	0	Y	Y	Y	Y	0	0	-
C17orf57	R236*	G->A	0	1	Y	Y	Y	Y	0	0	-
C17orf57	G->A	K433*	1	1	Y	Y	Y	Y	1	1	Y
KLK14	A->G	C->T	0	1	N	N	Y	Y	1	1	-
LILRA3	E54*	C->G	1	0	Y	Y	Y	Y	1	1	N
C20orf71	S109*	C165*	1	0	Y	Y	Y	Y	0	0	-
SLC17A9	G->T	T->C	1	0	Y	N	Y	N	1	1	N
FLG	R501*	R501*	1	0	Y	Y	Y	Y	0	0	-
CFHR2	E199*	E199*	1	0	Y	Y	Y	Y	0	0	-
NT5C1B	S118*	S118*	1	0	Y	Y	Y	Y	0	0	-
SULT1C3	W36*	W36*	1	0	Y	Y	Y	Y	0	0	-
THSD7B	C->T	C->T	1	0	Y	Y	Y	Y	0	0	-
DPP4	T->C	T->C	2	0	Y	Y	Y	Y	0	0	-

Table 2.6: List of Autosomal Genes with Rare Complete Knockouts (Continued)

Gene	Variant 1 Annotation	Variant 2 Annotation	# Cases	# Controls	Variant 1 found in dbSNP?	Variant 1 found in NHLBI ESP?	Variant 2 found in dbSNP?	Variant 2 found in NHLBI ESP?	# Individuals validated for Variant 1	# Individuals validated for Variant 2	*In trans?*
SLC22A14	A->C	A->C	2	0	Y	Y	Y	Y	0	0	-
TGM4	W269*	W269*	1	0	Y	Y	Y	Y	0	0	-
KNG1	R412*	R412*	1	0	N	N	N	N	0	0	-
ATP13A5	Q355*	Q355*	1	0	Y	Y	Y	Y	0	0	-
UGT2A1	Y192*	Y192*	1	0	Y	Y	Y	Y	0	0	-
MICB	R193*	R193*	1	0	Y	Y	Y	Y	0	0	-
LPA	C->T	C->T	1	0	Y	Y	Y	Y	0	0	-
ABCB5	G->C	G->C	2	0	Y	Y	Y	Y	0	0	-
CCL26	R44*	R44*	1	0	Y	Y	Y	Y	0	0	-
ACTR3C	Q121*	Q121*	1	0	N	N	N	N	0	0	-
PLAT	R561*	R561*	1	0	Y	Y	Y	Y	0	0	-
PKHD1L1	G->C	G->C	2	0	Y	Y	Y	Y	0	0	-
OR1J1	C235*	C235*	1	0	Y	Y	Y	Y	0	0	-
PTCHD3	R476*	R476*	1	0	Y	Y	Y	Y	0	0	-

Table 2.6: List of Autosomal Genes with Rare Complete Knockouts (Continued)

Gene	Variant 1 Annotation	Variant 2 Annotation	# Cases	# Controls	Variant 1 found in dbSNP?	Variant 1 found in NHLBI ESP?	Variant 2 found in dbSNP?	Variant 2 found in NHLBI ESP?	# Individuals validated for Variant 1	# Individuals validated for Variant 2	*In trans?*
CYP2C18	Y68*	Y68*	1	0	Y	Y	Y	Y	0	0	-
OR8K3	Q260*	Q260*	1	0	Y	Y	Y	Y	0	0	-
PZP	Q598*	Q598*	1	0	Y	N	Y	N	0	0	-
SPERT	C71*	C71*	1	0	Y	Y	Y	Y	0	0	-
CLYBL	R225*	R225*	1	0	Y	Y	Y	Y	0	0	-
OR4M2	R128*	R128*	1	0	Y	Y	Y	Y	0	0	-
SPTBN5	R1848*	R1848*	2	0	Y	Y	Y	Y	0	0	-
LRRC29	W6*	W6*	1	0	Y	Y	Y	Y	0	0	-
KRTAP4-8	C30*	C30*	1	0	Y	Y	Y	Y	0	0	-
KRT31	C->T	C->T	1	0	Y	Y	Y	Y	0	0	-
HAP1	S616*	S616*	1	0	Y	Y	Y	Y	0	0	-
C17orf57	R236*	R236*	1	0	Y	Y	Y	Y	0	0	-
C17orf57	G->A	G->A	1	0	Y	Y	Y	Y	0	0	-
ENDOV	A->G	A->G	1	0	Y	Y	Y	Y	0	0	-

Table 2.6: List of Autosomal Genes with Rare Complete Knockouts (Continued)

Gene	Variant 1 Annotation	Variant 2 Annotation	# Cases	# Controls	Variant 1 found in dbSNP?	Variant 1 found in NHLBI ESP?	Variant 2 found in dbSNP?	Variant 2 found in NHLBI ESP?	# Individuals validated for Variant 1	# Individuals validated for Variant 2	*In trans?*
LILRA3	E54*	E54*	1	0	Y	Y	Y	Y	0	0	-
LILRA3	C->G	C->G	1	0	Y	Y	Y	Y	0	0	-
USH2A	Y4238*	W2075*	1	0	N	N	N	N	0	0	Y
ZAN	G->A	G->A	1	0	Y	Y	Y	Y	0	0	Y

Table 2.7: List of X-chromosome Genes with Rare Complete Knockouts

Gene	Chr	Variant Position	Variant Annotation	# Male cases	# Male controls	# Female cases	# Female controls	Variant found in dbSNP?	Variant found in NHLBI ESP?	# Validated males
GYG2	X	2795295	Q431*	1	0	0	0	N	N	1
ARSH	X	2931214	G->T	1	0	0	0	N	N	0
FAM9C	X	13061246	C->A	1	0	0	0	N	N	0
PIR	X	15415636	Q210*	1	0	0	0	N	N	0
BEND2	X	18221904	Y208*	1	0	0	0	N	N	1
MAP3K15	X	19387332	R1136*	3	2	0	0	Y	Y	0
MAP3K15	X	19389113	R1122*	0	0	1	0	Y	Y	0
MAP3K15	X	19416381	R677*	0	1	0	0	Y	N	0
MAP3K15	X	19433376	C->A	0	1	0	0	Y	Y	0
MAP3K15	X	19482459	C197*	1	0	0	0	N	N	1
KLHL34	X	21675102	Q269*	0	1	0	0	Y	Y	0
PRDX4	X	23700511	A->G	1	0	0	0	N	N	0

Table 2.7: List of X-chromosome Genes with Rare Complete Knockouts (Continued)

Gene	Chr	Variant Position	Variant Annotation	# Male cases	# Male controls	# Female cases	# Female controls	Variant found in dbSNP?	Variant found in NHLBI ESP?	# Validated males
GYG2	X	2795295	Q431*	1	0	0	0	N	N	1
ARSH	X	2931214	G->T	1	0	0	0	N	N	0
FAM9C	X	13061246	C->A	1	0	0	0	N	N	0
PIR	X	15415636	Q210*	1	0	0	0	N	N	0
BEND2	X	18221904	Y208*	1	0	0	0	N	N	1
MAP3K15	X	19387332	R1136*	3	2	0	0	Y	Y	0
MAP3K15	X	19389113	R1122*	0	0	1	0	Y	Y	0
MAP3K15	X	19416381	R677*	0	1	0	0	Y	N	0
MAP3K15	X	19433376	C->A	0	1	0	0	Y	Y	0
MAP3K15	X	19482459	C197*	1	0	0	0	N	N	1
KLHL34	X	21675102	Q269*	0	1	0	0	Y	Y	0
PRDX4	X	23700511	A->G	1	0	0	0	N	N	0
FTHL17	X	31089629	E148*	0	1	0	0	Y	Y	0

Table 2.7: List of X-chromosome Genes with Rare Complete Knockouts (Continued)

Gene	Chr	Variant Position	Variant Annotation	# Male cases	# Male controls	# Female cases	# Female controls	Variant found in dbSNP?	Variant found in NHLBI ESP?	# Validated males
DUSP21	X	44703448	Q24*	0	1	0	0	N	N	1
ZNF157	X	47272856	R462*	1	0	0	0	N	N	1
SSX1	X	48125772	Q173*	0	1	0	0	N	N	1
GAGE8	X	49236822	G->T	1	0	0	0	N	Y	Failed primer design
ITIH5L	X	54815083	C->T	1	0	0	0	N	N	1
PFKFB1	X	54987256	C->T	1	0	0	0	N	N	Failed primer design
FAAH2	X	57475022	E432*	3	1	1	0	Y	Y	2
MTMR8	X	63565059	T->G	1	0	0	0	N	N	1
VSIG4	X	65242332	R325*	2	0	0	0	Y	Y	1
DGAT2L6	X	69420288	R151*	1	0	0	0	Y	Y	1
DGAT2L6	X	69421907	Q214*	1	0	0	0	N	Y	1
CXCR3	X	70837390	Q25*	1	0	0	0	Y	Y	0
KIAA2022	X	73959981	Q1471*	1	0	0	0	N	N	1

Table 2.7: List of X-chromosome Genes with Rare Complete Knockouts (Continued)

Gene	Chr	Variant Position	Variant Annotation	# Male cases	# Male controls	# Female cases	# Female controls	Variant found in dbSNP?	Variant found in NHLBI ESP?	# Validated males
SATL1	X	84349207	Y601*	1	0	0	0	Y	Y	1
PCDH11X	X	91137905	G->A	1	0	0	0	Y	Y	0
SRPX2	X	99921931	G->A	1	0	0	0	N	N	1
DRP2	X	100503119	E432*	1	0	0	0	N	N	1
GLRA4	X	102979868	R54*	1	0	0	0	Y	Y	0
VSIG1	X	107316600	G->A	1	0	0	0	N	N	1
GUCY2F	X	108673541	E596*	1	1	0	0	N	N	1
ZCCHC16	X	111698851	R299*	1	0	0	0	N	N	1
LUZP4	X	114540908	R161*	1	0	0	0	N	N	1
LUZP4	X	114540914	R163*	1	0	0	0	N	N	1
KIAA1210	X	118220508	C->T	1	0	0	0	Y	N	1
SLC25A43	X	118544221	R196*	1	0	0	0	Y	N	1
SLC25A43	X	118586969	R323*	1	0	0	0	Y	Y	1

Table 2.7: List of X-chromosome Genes with Rare Complete Knockouts (Continued)

Gene	Chr	Variant Position	Variant Annotation	# Male cases	# Male controls	# Female cases	# Female controls	Variant found in dbSNP?	Variant found in NHLBI ESP?	# Validated males
ATP1B4	X	119510234	G->C	1	0	0	0	Y	N	0
ZNF75D	X	134421668	Q312*	1	2	0	0	Y	Y	0
MMGT1	X	135047209	R124*	0	1	0	0	N	Y	0
MAP7D3	X	135328223	C->A	0	1	0	0	N	N	1
GPR112	X	135428491	Q876*	1	0	0	0	N	N	1
MCF2	X	138664615	S857*	1	0	0	0	Y	Y	0
MAGEC3	X	140969250	Q193*	1	0	0	0	Y	Y	1
AFF2	X	147744095	Q283*	1	0	0	0	N	N	1
MAMLD1	X	149638986	R381*	0	1	0	0	N	N	1
MECP2	X	153295832	E495*	1	0	0	0	N	N	1
TMLHE	X	154743928	T->C	1	0	0	0	N	N	1
ARSF	X	2994704	R93*	1	0	0	0	Y	Y	0
FAM9C	X	13058887	Q107*	0	1	0	0	N	Y	0

Table 2.7: List of X-chromosome Genes with Rare Complete Knockouts (Continued)

Gene	Chr	Variant Position	Variant Annotation	# Male cases	# Male controls	# Female cases	# Female controls	Variant found in dbSNP?	Variant found in NHLBI ESP?	# Validated males
DMD	X	31196048	C->T	1	0	0	0	Y	Y	0
SSX1	X	48116655	A->T	1	0	0	0	N	Y	0
ITIH6	X	54777550	E1206*	0	1	0	0	Y	Y	0
VSIG4	X	65244866	C->T	0	1	0	0	N	N	0
DGAT2L6	X	69397495	W21*	1	0	0	0	N	N	0
P2RY4	X	69479172	W101*	0	1	0	0	Y	Y	0
ERCC6L	X	71424992	E1086*	0	1	0	0	N	N	0
ZCCHC13	X	73524146	W15*	1	0	0	0	Y	Y	0
SATL1	X	84362650	W255*	2	2	0	0	Y	Y	0
SYTL4	X	99946207	C->T	0	1	0	0	N	N	0
TMEM31	X	102968546	Q43*	0	1	0	0	N	N	0
GLRA4	X	102979868	R54*	1	1	0	0	Y	Y	0
RNF128	X	105937343	Y37*	1	0	0	0	N	N	0

Table 2.7: List of X-chromosome Genes with Rare Complete Knockouts (Continued)

Gene	Chr	Variant Position	Variant Annotation	# Male cases	# Male controls	# Female cases	# Female controls	Variant found in dbSNP?	Variant found in NHLBI ESP?	# Validated males
TEX13B	X	107224898	C->T	1	1	0	0	Y	Y	0
SLC25A43	X	118540640	R165*	0	1	0	0	Y	Y	0
OR13H1	X	130678702	R219*	1	0	0	0	N	Y	0
CT45A5	X	134947910	R139*	1	0	0	0	Y	Y	0
MAP7D3	X	135312997	C->G	1	0	0	0	N	Y	0
CDR1	X	139865915	W206*	0	1	0	0	N	N	0
MAGEC3	X	140969250	Q193*	1	0	0	0	Y	Y	0
HAUS7	X	152720388	R362*	1	0	0	0	N	N	0

Similar Enrichment of Rare Complete Knockouts Observed on the X-chromosome

Given the gender bias in ASD, with roughly 4 times as many affected males than females [24], we asked analogously whether rare gene knockouts outside the pseudo-autosomal regions on the X-chromosome (arising from hemizygous LoFs in males) were enriched in male cases versus male controls. To further increase the sample sizes, we included the male probands and their unaffected fathers from the trios and quartets. The nucleotide diversity on the X-chromosome is estimated to be between half to three-quarters that of the autosomes and deleterious LoF variants on the X-chromosome are under stronger negative selection given the smaller effective population size and constant exposure in hemizygous males [25]. To match the baseline knockout rate to the autosomes, where we examined variants with ≤5% minor allele frequency (MAF) and therefore ≤0.25% homozygosity, we examined LoF variants with population frequency (assessed in female control samples) of ≤0.25%. On average, we observed less than 1 such rare LoF variant on the X-chromosome in both males and females (Table 2.8).

Table 2.8: Population Distribution of LoF variants on the X-chromosome

The average number of rare (≤0.25%) and common (>0.25%) LoFs outside the pseudo-autosomal regions on the X-chromosome in males, as well as the average number of rare and common LoFs outside the pseudo-autosomal regions on the X-chromosome in females is shown.

	Average number of hemizygous variants in males	Number of unique genes with a hemizygous variant in males in the datasets	Average number of heterozygous and homozygous variants in females	Number of unique genes with a variant in females in the datasets
Rare (≤0.25%) LoFs	**0.02 variant per individual**	**28 genes**	**0.04 variant per individual**	**41 genes**
Common (>0.25%) LoFs	**0.23 variant per individual**	**11 genes**	**0.21 variant per individual**	**13 genes**

Similar to the autosomes, we observed a significant enrichment of rare hemizygous LoFs in male cases (Table 2.9), with 88 such events observed – 60 of them were found in male cases and 28 of them were found in male controls (OR = 1.5, 95% CI = [1.1, 2.0], one-sided hypergeometric test $P = 0.034$). No enrichment was seen in the internal controls of this comparison - rare hemizygous synonymous variants were not enriched in male cases compared to male controls (OR = 1.0, 95% CI = [0.9, 1.1]), indicating the observed enrichment is specific to rare complete knockouts on the X-chromosome in male ASD cases. Based on the difference between cases and controls, we further estimate another 1.7% contribution by rare complete knockouts on the X-chromosome in male cases. In addition, we found 2 of 170 female cases bearing a rare complete knockout on the X-chromosome and 0 of 452 female controls. As with the autosomes, we attempted validation for 44 of 50 rare X-chromosome LoF variants and all 44 validated.

We screened the list of rare complete knockouts observed on the autosomes and X-chromosome for instances where a knockout was observed only in cases and not in any of the controls (Table 2.10) and performed a screen for enrichment of pathways and microRNA targets using WebGestalt [26]. The top pathway ("Complement and coagulation cascades") was driven by 2 genes (*KNG1* and *PLAT*; corrected $P = 0.0027$). Scanning predicted targets of microRNAs, we found one (*mir-328*) predicted to target 3 genes from the list (*HAP1*, *AFF2* and *MECP2*; corrected $P = 0.0013$; Table 2.11). Additional siblings (affected = 30, unaffected = 17) were available for 31 probands who were genotyped to examine segregation of a proposed recessive model (Table 2.12). We observed 25 (expected 20) instances where segregation was consistent with a fully penetrant recessive model, including 4 genes with rare complete knockouts (*PTH2R*,

MECP2, *VSIG1* and *ZCCHC16*) observed in cases only and not in a single control in any wave of our study.

Gender and IQ

It has been shown that the male gender bias is stronger in high-functioning ASD cases, and the gender bias is reduced for syndromic cases [27]. We found that there was a higher rate of rare complete knockouts in females (5.4%) compared to males (4%). Although 16% of the cases sequenced were female, 25% of the cases harboring rare complete knockouts were female (OR = 1.7, 95% CI = [1.3, 2.1], one-sided Fisher's $P = 0.076$). While not statistically significant, this trend is similar to previous observations that *de novo* CNVs and SNVs show a higher fraction of female cases with such events [4,5,8] and consistent with the model that females need a higher dose of genetic risk to manifest a diagnosis of ASD. We also observed a trend in IQ scores from 18 of these cases with rare complete knockouts to another 133 cases (mean Z-score = -0.26 in probands with rare complete knockouts versus 0.035 in other cases), but it was not statistically significant (one-sided Wilcox $P = 0.11$).

Table 2.9: Number of Rare LoF and Synonymous Variants on the X-chromosome

The number of rare hemizygous LoF and synonymous variants outside the pseudo-autosomal regions on the X-chromosome in males, as well as the number of rare heterozygous LoF and synonymous variants in females are shown, together with the respective odds ratios.

	Rare hemizygous / heterozygous LoF variants	Rare hemizygous / heterozygous synonymous variants
Hemizygous LoFs in males (N = 2,144)		
Cases (N = 1,245)	60 events	2,114 events
Controls (N = 899)	28 events	1,516 events
OR [95% CI]	1.5 [1.1, 2.0]	1.0 [0.9, 1.1]
Heterozygous LoFs in females (N = 622)		
Cases (N = 170)	21 events	641 events
Controls (N = 452)	56 events	1,256 events
OR [95% CI]	1.0 [0.5, 1.5]	1.4 [1.2, 1.6]
	Rare homozygous LoF variants	**Rare homozygous synonymous variants**
Homozygous LoFs in females (N = 622)		
Cases (N = 170)	2 events	5 events
Controls (N = 452)	0 events	0 events
OR [95% CI]	-	-

Table 2.10: List of Rare Complete Knockouts on Autosomes and X-chromosome Found in Cases Only

A summary of the list of genes with rare complete knockouts observed only in the cases and not in controls - genes found to be involved in known diseases have been marked with "*", and genes found in CNVs regions previously implicated in ASD risk have been marked with "+".

Gene	Chr	# Cases	# Controls	Expressed in the brain?
DGAT2L6	X	3	0	No
SLC22A14	3	2	0	No
LUZP4	X	2	0	No
MAGEC3	X	2	0	Unknown
CFHR2	1	1	0	Yes
USH2A*	1	1	0	No
PTH2R	2	1	0	Yes
KNG1	3	1	0	No
TGM4	3	1	0	No
AGXT2	5	1	0	No
KIAA1919	6	1	0	Yes
MICB	6	1	0	Yes
ACTR3C	7	1	0	Unknown
LRRC69	8	1	0	Unknown
PLAT	8	1	0	Yes
CYP2C18	10	1	0	Yes

Table 2.10: List of Rare Complete Knockouts on Autosomes and X-chromosome Found in Cases Only (Continued)

A summary of the list of genes with rare complete knockouts observed only in the cases and not in controls - genes found to be involved in known diseases have been marked with "*", and genes found in CNVs regions previously implicated in ASD risk have been marked with "$^{+}$".

C12orf64	12	1	0	Unknown
PZP	12	1	0	Unknown
LRRC29	16	1	0	No
DBF4B	17	1	0	Unknown
HAP1	17	1	0	Yes
AFF2*	X	1	0	Yes
ARSF	X	1	0	Yes
ARSH	X	1	0	Unknown
ATP1B4	X	1	0	No
BEND2	X	1	0	No
CT45A5	X	1	0	Yes
CXCR3	X	1	0	Yes
DMD	X	1	0	Yes
DRP2	X	1	0	Yes
GPR112	X	1	0	Unknown
GYG2	X	1	0	Yes
HAUS7	X	1	0	Yes
ITIH5L	X	1	0	No

Table 2.10: List of Rare Complete Knockouts on Autosomes and X-chromosome Found in Cases Only (Continued)

A summary of the list of genes with rare complete knockouts observed only in the cases and not in controls - genes found to be involved in known diseases have been marked with "*", and genes found in CNVs regions previously implicated in ASD risk have been marked with "+".

KIAA1210	X	1	0	Unknown
KIAA2022*	X	1	0	Unknown
MCF2	X	1	0	Yes
MECP2*	X	1	0	Yes
MTMR8	X	1	0	Yes
PCDH11X+	X	1	0	Yes
PIR	X	1	0	Yes
PRDX4	X	1	0	Yes
RNF128	X	1	0	Yes
SRPX2*	X	1	0	Yes
TMLHE+	X	1	0	Yes
VSIG1	X	1	0	Yes
ZCCHC13	X	1	0	No
ZCCHC16	X	1	0	No
ZNF157	X	1	0	Yes

Table 2.11: List of microRNAs and the Enrichment of Targets in the Gene List Found in Table 2.10

MicroRNA	Raw P-value	Corrected P-value
MIR-328	6.28e-05	0.0013
MIR-361	0.0031	0.0163
MIR-504	0.0030	0.0163
MIR-124A	0.0020	0.0163
MIR-194	0.0049	0.0175
MIR-452	0.0050	0.0175
MIR-143	0.0090	0.0270
MIR-494	0.0105	0.0276
MIR-518A-2	0.0166	0.0349
MIR-369-3P	0.0166	0.0349

Table 2.12: Segregation Patterns within Multiplex Families

Gene	Status	Gender	Affected_status	MZ/DZ	Genotype1	Genotype2	Validated?	In *trans*?	Recessive inheritance?
C1orf127	Proband	F	Affected		AA	-	Y	-	N
	Sibling	F	Unaffected		AA	-			
	Sibling	M	Affected		GA	-			
	Dad	M	Unaffected		AA	-			
	Mum	F	Unaffected		GA	-			
PRAMEF2	Proband	M	Affected		T T	-	Y	-	N
	Sibling	M	Affected		GG	-			
	Sibling	M	Unaffected		GT	-			
	Dad	M	Unaffected		GT	-			
	Mum	F	Unaffected		GG	-			
C1orf168	Proband	M	Affected		TC	TA	Y	N	-
	Sibling	F	Affected		TT	TT			
	Dad	M	Unaffected		TT	TT			
	Mum	F	Unaffected		TC	TA			

Table 2.12: Segregation Patterns within Multiplex Families (Continued)

Gene	Status	Gender	Affected_status	MZ/DZ	Genotype1	Genotype2	Validated?	In *trans*?	Recessive inheritance?
VWA3B	Proband	M	Affected		TT	-	Y	-	N
	Sibling	M	Affected		GT	-			
	Dad	M	Unaffected		TT	-			
	Mum	F	Unaffected		GT	-			
VWA3B	Proband	M	Affected		GT	AA	N	-	-
	Sibling	M	Affected		GT	AA			
	Dad	M	Unaffected		GT	AA			
	Mum	F	Unaffected		GG	AA			
PTH2R	Proband	M	Affected		AA	-	Y	-	Y
	Sibling	M	Affected		AA	-			
	Sibling	M	Unaffected		CA	-			
	Sibling	M	Unaffected		CC	-			
	Dad	M	Unaffected		CA	-			
	Mum	F	Unaffected		CA	-			
DPP4	Proband	M	Affected		CC	-	Y	-	Y
	Sibling	M	Affected		CC	-			

Table 2.12: Segregation Patterns within Multiplex Families (Continued)

Gene	Status	Gender	Affected_status	MZ/DZ	Genotype1	Genotype2	Validated?	In *trans*?	Recessive inheritance?
	Sibling	F	Unaffected		TT	-			
	Dad	M	Unaffected		TC	-			
	Mum	F	Unaffected		TC	-			
DPP4	Proband	F	Affected		CC	-	Y	-	N
	Sibling	F	Unaffected		TC	-			
	Sibling	M	Affected		CC	-			
	Dad	M	Unaffected		CC	-			
	Mum	F	Unaffected		TC	-			
IFIH1	Proband	M	Affected		GG	-	Y	-	N
	Sibling	M	Affected		CC	-			
	Sibling	M	Unaffected		CC	-			
	Dad	M	Unaffected		CG	-			
	Mum	F	Unaffected		CG	-			
UGT2A1	Proband	M	Affected		TT	-	Y	-	N
	Sibling	M	Affected		AA	-			
	Sibling	M	Unaffected		AT	-			

Table 2.12: Segregation Patterns within Multiplex Families (Continued)

Gene	Status	Gender	Affected_status	MZ/DZ	Genotype1	Genotype2	Validated?	In *trans*?	Recessive inheritance?
	Dad	M	Unaffected		AT	-			
	Mum	F	Unaffected		AT	-			
WDR17	Proband	F	Affected		AA	AA	N	-	-
	Sibling	F	Affected		AA	AA			
	Dad	M	Unaffected		AA	AA			
	Mum	F	Unaffected		AA	AA			
KIAA1919	Proband	F	Affected		AA	-	Y	-	N
	Sibling	F	Affected		TT	-			
	Sibling	M	Unaffected		TA	-			
	Dad	M	Unaffected		TA	-			
	Mum	F	Unaffected		TA	-			
TAAR2	Proband	F	Affected		TT	-	Y	-	Y
	Sibling	F	Affected		TT	-			
	Dad	M	Unaffected		CT	-			
	Mum	F	Unaffected		CT	-			
UNC93A	Proband	M	Affected		CC	-	Y	-	N

Table 2.12: Segregation Patterns within Multiplex Families (Continued)

Gene	Status	Gender	Affected_status	MZ/DZ	Genotype1	Genotype2	Validated?	In *trans*?	Recessive inheritance?
	Sibling	M	Affected		GC	-			
	Sibling	M	Affected		GC	-			
	Dad	M	Unaffected		GC	-			
	Mum	F	Unaffected		GC	-			
RNF32	Proband	M	Affected		CC	-	Y	-	N
	Sibling	M	Affected		TT	-			
	Sibling	F	Affected		CC	-			
	Dad	M	Unaffected		TC	-			
	Mum	F	Unaffected		TC	-			
RNF32	Proband	F	Affected		CC	-	Y	-	N
	Sibling	F	Affected		TT	-			
	Dad	M	Unaffected		TC	-			
	Mum	F	Unaffected		TC	-			
INTS8	Proband	F	Affected		AA	AA	N	-	-
	Sibling	F	Affected		AA	AA			
	Dad	M	Unaffected		AA	AA			

Table 2.12: Segregation Patterns within Multiplex Families (Continued)

Gene	Status	Gender	Affected_status	MZ/DZ	Genotype1	Genotype2	Validated?	In *trans*?	Recessive inheritance?
	Mum	F	Unaffected		AA	AA			
OR10V1	Proband	M	Affected		AA	-	Y	-	N
	Sibling	F	Affected		AA	-			
	Dad	M	Unaffected		AA	-			
	Mum	F	Unaffected		AA	-			
SLC22A25	Proband	M	Affected		TC	CT	Y	Y	Y
	Sibling	M	Affected		TC	CT			
	Dad	M	Unaffected		TC	TT			
	Mum	F	Unaffected		TT	CT			
RAD52	Proband	F	Affected		AC	-	Y	-	Y
	Sibling	F	Affected		CC	-			
	Dad	M	Unaffected		AC	-			
	Mum	F	Unaffected		AC	-			
RAD52	Proband	M	Affected		AC	GT	Y	-	N
	Sibling	M	Affected		CC	GG			
	Sibling	F	Unaffected		CC	GG			

Table 2.12: Segregation Patterns within Multiplex Families (Continued)

Gene	Status	Gender	Affected_status	MZ/DZ	Genotype1	Genotype2	Validated?	In *trans*?	Recessive inheritance?
	Dad	M	Unaffected		AC	TT			
	Mum	F	Unaffected		AC	GG			
KRT83	Proband	M	Affected		AA	-	Y	-	N
	Sibling	M	Unaffected		CA	-			
	Sibling	M	Unaffected		AA	-			
	Dad	M	Unaffected		AA	-			
	Mum	F	Unaffected		CA	-			
OAS3	Proband	F	Affected		CT	CT	Y	Y	N
	Sibling	F	Affected		CT	CC			
	Dad	M	Unaffected		CT	CC			
	Mum	F	Unaffected		CC	CT			
C14orf105	Proband	M	Affected		AA	-	Y	-	N
	Sibling	M	Affected		GA	-			
	Dad	M	Unaffected		GA	-			
	Mum	F	Unaffected		GA	-			
ABCC12	Proband	F	Affected		TT	-	Y	-	N

Table 2.12: Segregation Patterns within Multiplex Families (Continued)

Gene	Status	Gender	Affected_status	MZ/DZ	Genotype1	Genotype2	Validated?	In *trans*?	Recessive inheritance?
	Sibling	M	Affected		CT	-			
	Sibling	M	Unaffected		CT	-			
	Dad	M	Unaffected		CT	-			
	Mum	F	Unaffected		CT	-			
C17orf57	Proband	M	Affected		GA	AT	Y	Y	Y
	Sibling	F	Unaffected		GA	AA			
	Dad	M	Unaffected		GA	AA			
	Mum	F	Unaffected		GG	AT			
ABCA10	Proband	M	Affected		AA	-	Y	-	Y
	Sibling	F	Affected	MZ	AA	-			
	Sibling	F	Affected	MZ	AA	-			
	Dad	M	Unaffected		GA	-			
	Mum	F	Unaffected		GA	-			
ABCA10	Proband	M	Affected		AA	-	Y	-	N
	Sibling	M	Affected		GA	-			
	Sibling	M	Unaffected		GG	-			

Table 2.12: Segregation Patterns within Multiplex Families (Continued)

Gene	Status	Gender	Affected_status	MZ/DZ	Genotype1	Genotype2	Validated?	In *trans*?	Recessive inheritance?
	Dad	M	Unaffected		GA	-			
	Mum	F	Unaffected		GA	-			
ABCA10	Proband	M	Affected		AA	-	Y	-	Y
	Sibling	M	Affected		AA	-			
	Sibling	M	Unaffected		GA	-			
	Dad	M	Unaffected		GA	-			
	Mum	F	Unaffected		GA	-			
ABCA10	Proband	M	Affected		AA	-	Y	-	N
	Sibling	M	Affected		GA	-			
	Sibling	F	Affected		GA	-			
	Dad	M	Unaffected		GA	-			
	Mum	F	Unaffected		GA	-			
FLJ35220	Proband	M	Affected		GG	-	Y	-	N
	Sibling	M	Affected		AG	-			
	Dad	M	Unaffected		GG	-			
	Mum	F	Unaffected		AG	-			

Table 2.12: Segregation Patterns within Multiplex Families (Continued)

Gene	Status	Gender	Affected_status	MZ/DZ	Genotype1	Genotype2	Validated?	In *trans*?	Recessive inheritance?
STARD6	Proband	M	Affected		AA	-	Y	-	N
	Sibling	M	Affected		GA	-			
	Dad	M	Unaffected		GA	-			
	Mum	F	Unaffected		AA	-			
STARD6	Proband	F	Affected		AA	-	Y	-	N
	Sibling	F	Unaffected		AA	-			
	Sibling	M	Affected		GG	-			
	Dad	M	Unaffected		GA	-			
	Mum	F	Unaffected		GA	-			
STARD6	Proband	M	Affected		AA	-	Y	-	Y
	Sibling	M	Unaffected		GA	-			
	Sibling	M	Affected		AA	-			
	Dad	M	Unaffected		GA	-			
	Mum	F	Unaffected		GA	-			
ZNF77	Proband	M	Affected		AA	-	Y	-	N
	Sibling	M	Affected		GA	-			

Table 2.12: Segregation Patterns within Multiplex Families (Continued)

Gene	Status	Gender	Affected_status	MZ/DZ	Genotype1	Genotype2	Validated?	In *trans*?	Recessive inheritance?
	Dad	M	Unaffected		GA	-			
	Mum	F	Unaffected		GA	-			
UNC13A	Proband	M	Affected		AA	-	Y	-	N
	Sibling	M	Affected		GA	-			
	Dad	M	Unaffected		GA	-			
	Mum	F	Unaffected		AA	-			
LILRA3	Proband	M	Affected		CA	CG	Y	N	N
	Sibling	M	Affected		CA	CG			
	Sibling	M	Unaffected		CA	CG			
	Dad	M	Unaffected		CC	CC			
	Mum	F	Unaffected		CA	CG			
SLC17A9	Proband	F	Affected		GT	TC	Y	N	-
	Sibling	F	Unaffected		GG	TT			
	Sibling	F	Unaffected		GG	TT			
	Sibling	M	Affected		GT	TC			
	Dad	M	Unaffected		GG	TT			

Table 2.12: Segregation Patterns within Multiplex Families (Continued)

Gene	Status	Gender	Affected_status	MZ/DZ	Genotype1	Genotype2	Validated?	In *trans*?	Recessive inheritance?
	Mum	F	Unaffected		GT	TC			
MAP3K15	Proband	M	Affected		TT	-	Y	-	N
	Sibling	M	Unaffected		TT	-			
	Dad	M	Unaffected		GG	-			
	Mum	F	Unaffected		GT	-			
ZNF157	Proband	M	Affected		TT	-	Y	-	N
	Sibling	M	Affected		CC	-			
	Dad	M	Unaffected		CC	-			
	Mum	F	Unaffected		CT	-			
MTMR8	Proband	M	Affected		GG	-	Y	-	N
	Sibling	M	Affected		TT	-			
	Sibling	M	Unaffected		TT	-			
	Sibling	F	Unaffected		TT	-			
	Sibling	F	Unaffected		TT	-			
	Dad	M	Unaffected		TT	-			
	Mum	F	Unaffected		TG	-			

Table 2.12: Segregation Patterns within Multiplex Families (Continued)

Gene	Status	Gender	Affected_status	MZ/DZ	Genotype1	Genotype2	Validated?	In *trans*?	Recessive inheritance?
VSIG4	Proband	M	Affected	MZ	AA	-	Y	-	Y
	Sibling	M	Affected	MZ	AA	-			
	Sibling	M	Unaffected		TT	-			
	Dad	M	Unaffected		TT	-			
	Mum	F	Unaffected		TA	-			
VSIG1	Proband	M	Affected		AA	-	Y	-	Y
	Sibling	M	Affected		AA	-			
	Sibling	M	Unaffected		GG	-			
	Dad	M	Unaffected		GG	-			
	Mum	F	Unaffected		GA	-			
ZCCHC16	Proband	M	Affected		TT	-	Y	-	Y
	Sibling	M	Affected		TT	-			
	Sibling	F	Unaffected		CC	-			
	Dad	M	Unaffected		CC	-			
	Mum	F	Unaffected		CT	-			
LUZP4	Proband	M	Affected		TT	-	Y	-	N

Table 2.12: Segregation Patterns within Multiplex Families (Continued)

Gene	Status	Gender	Affected_status	MZ/DZ	Genotype1	Genotype2	Validated?	In *trans*?	Recessive inheritance?
	Sibling	M	Affected	MZ	CC	-			
	Sibling	M	Affected	MZ	CC	-			
	Dad	M	Unaffected		CC	-			
	Mum	F	Unaffected		CT	-			
SLC25A43	Proband	M	Affected		TT	-	Y	-	N
	Sibling	M	Unaffected		TT	-			
	Dad	M	Unaffected		CC	-			
	Mum	F	Unaffected		CT	-			
MECP2	Proband	M	Affected		AA	-	Y	-	Y
	Sibling	M	Affected		AA	-			
	Dad	M	Unaffected		CC	-			
	Mum	F	Unaffected		AC	-			

DISCUSSION

As shown previously, *de novo* copy number variants (CNVs) are extremely rare events in a control population and they occur at 1-2% in controls. Given the rarity of such events, discovery of a global enrichment of these *de novo* CNVs at a much higher rate of 6-8% in ASD individuals suggested a 6% contribution to ASD by these *de novo* CNVs [2,4,5]. This highlighted the significance of such events as risk factors for ASD and subsequent association and replication studies of such events with larger sample sizes pinpointed to specific *de novo* CNVs that have since been significantly associated with ASD, such as deletions and duplications on chromosome 16p11.2 [3].

Similar to the *de novo* CNV studies, as well as emerging *de novo* SNV studies, we observed that rare complete knockouts in the human exome are found in only 3% of a control population, but are present at a 2-fold enrichment in ASD cases. Given that these rare complete knockouts are not found in a single gene but, like the *de novo* CNVs and SNVs, are distributed across many different genes, these events would have been missed through previous association or linkage studies. As with any genetic screen, population stratification can confound these results. However, the samples selected for sequencing were of European ancestry and individually matched in case-control pairs based on principal component analyses and selected from a much larger pool of potential samples. Owing to occasional sample failure, ultimately 88% of the final samples were matched one-to-one for ancestry and a similar 2-fold enrichment was observed in the subset of matched cases and controls for the rare complete knockouts (49 events in cases versus 25 events in controls, OR = 2, 95% CI = [1.5, 2.5]).

Interestingly, we observed a 1.5-fold enrichment of hemizygous LoF variants on the X-chromosome in male cases compared to male controls, but did not observe a significant global

enrichment of heterozygous LoF variants on the X-chromosome in female cases compared to female controls. There are genes on the X-chromosome that can cause ASD-related disorders like Rett Syndrome in an X-linked dominant mode of inheritance such as *CDKL5* and *MECP2*. However, we found that while there is a significant 1.5-fold enrichment in hemizygous LoFs in male cases, we did not observe a significant enrichment in single-copy losses in female cases, consistent with the observation that we did not see an overall difference in single-copy (heterozygous) losses on the autosomes. Given that males have only a single copy of the X-chromosome and would be more susceptible to a complete knockout on the X-chromosome than females, these rare complete knockouts on the X-chromosome can also explain a small part of the male gender bias observed in ASD.

Candidate genes

Among our list of consolidated genes with rare complete knockouts that were observed only in cases, we discovered a known autosomal recessive gene in one of the probands from the trios – *Usher syndrome 2A* (*USH2A*), which has been reported to cause a known autosomal recessive disease Usher Syndrome Type II, characterized by mild to severe hearing loss and sometimes retinitis pigmentosa [28]. We found and confirmed the bilinieal inheritance of two previously unreported compound heterozygous nonsense mutations (W2075X and Y4238X) in *USH2A* from both parents. Clinical follow-up confirmed an Usher Syndrome Type II diagnosis – a potential confounder in the diagnosis of ASD [29].

When we cross-compared the list of genes harboring rare complete knockouts with previously published literature on *de novo* SNVs [7,8,9,10], we found 3 genes that were common between the rare complete knockouts and *de novo* SNVs – *IFIH1* (where a *de novo* missense

variant was found in a proband), *ABCC12* (where a *de novo* silent variant was found in a proband) and *PKHD1L1* (where a *de novo* upstream variant was found in a proband).

We further compared the list of X-chromosome genes with previously published CNVs and found that there are 2 genes that have been previously associated with rare CNVs. We found an affected male with a rare hemizygous splice variant (c.359-2T>C) in the *trimethyllysine hydroxylase, epsilon* protein – *TMLHE,* which is involved in the biosynthesis of carnitine [30]. Recently, *TMLHE* deficiency resulting in dysregulation of carnitine metabolism has also been proposed as a risk factor for ASD [31,32]. Another affected male was found to harbor a hemizygous splice variant (c.3034-1G>A) in the *protocadherin 11 X-linked* protein – *PCDH11X*. An inherited deletion in *PCDH11X*, as well as a *de novo* deletion in *PCDH11Y* was previously reported in a child with severe language delay, suggesting a potential role for *PCDH11X* in language development [33].

There were 3 genes with at least 2 male cases harboring rare complete knockouts on the X-chromosome and no controls were found to harbor rare complete knockouts in these genes (*SLC22A14*, *LUZP4*, *DGAT2L6*). In addition, among a list of genes known to be involved in intellectual disability [9], we found 4 genes from our list with rare complete knockouts in 4 male cases. One affected male has a nonsense variant Q283X in the *Fragile X E mental retardation syndrome protein* (*AFF2*), which causes non-syndromic mental retardation and this nonsense variant results in more than 80% of the protein to be truncated. Another male case has a nonsense variant resulting in Q1471X in an uncharacterized gene *KIAA2022* and mouse studies revealed that the protein is expressed in the developing brain and plays a role in neurite outgrowth [34]. A third male case has a splice variant c.961+1G>A in *Sushi-repeat containing protein, X-linked 2* (*SRPX2*), a protein that is found to be expressed in neurons. Mutations in *SRPX2* have been

reported to be associated with rolandic epilepsy with speech and cognition impairment [35] and *FOXP2*, a gene which is involved in speech and language disorders, has been shown to regulate *SRPX2* [36]. A fourth male with ASD harbored an E495X nonsense variant in *methyl CpG binding protein 2* (*MECP2*). Complete knockouts in *MECP2* are lethal in males and heterozygous LoFs in *MECP2* cause Rett Syndrome in females. Interestingly, the hemizygous nonsense mutation that was observed in this male case truncates only the last four amino acids of the *MECP2* protein and this potentially generates a protein product, which explains why the hemizygous LoF observed in this gene is viable in a male. Late-truncating mutations in *MECP2* have been reported to cause the Zappella variant of Rett Syndrome, which is a milder form of Rett Syndrome and autistic behavior is often observed in affected individuals [37].

Total Contribution to ASD From *de novo* and Inherited Factors

As described previously in various studies, there is an estimated 6% contribution to ASD risk from *de novo* CNVs [4,5,38]. Recent studies have estimated another 10% contribution to ASD risk by *de novo* SNVs [7,8,9,10]. In this study, we estimate a 3% contribution to ASD risk by rare complete knockouts on the autosomes and another 2% contribution by rare complete knockouts on the X-chromosome, resulting in another 5% contribution to ASD risk. Because a comparably reliable and validated set of insertion and deletion variants are not yet available across our entire dataset, we have not fully evaluated the contribution of frameshifts. Given that there is likely a similar number of frameshift mutations as single nucleotide LoF variants [8,19], the addition of frameshifts will likely increase this contribution further.

The global enrichment of rare complete knockouts in cases highlights the significance of such events in the overall genetic etiology of ASD. In addition, these events provide further

insight into the heritable component of ASD, which have not yet been accounted for by *de novo* CNVs and SNVs. However, many of these rare complete knockouts are distributed across many different genes. This agrees with our current understanding of ASD genetics to date: that this complex disorder follows a multigenic model where hundreds of genes are involved and that each individual gene accounts for a small fraction of ASD. Together with the ongoing *de novo* CNV and SNV studies, our study and that of another study in this issue [39], demonstrate convincing evidence of a rare recessive contribution to the heritability of ASD.

MATERIALS AND METHODS

The institutional review board of all participating institutions approved this study and written informed consent from all subjects was obtained. The datasets and detailed information for the samples have been deposited into dbGAP (accession ID: phs000298.v1.p1).

Exome capture and sequencing

Exome capture and sequencing at BI was performed as follows. Genomic DNA was sheared to 200-300 bp using a Covaris Acoustic Adaptor. Fragments were end-repaired, dA-tailed, and sequencing adaptor oligonucleotides ligated using reagents from New England BioLabs. Libraries were barcoded using the Illumina index read strategy, which uses six-base sequences within the adapter that are sequenced separately from the genomic DNA insert. The DNA library was subsequently enriched for sequences with 5' and 3' adapters by PCR amplication using with primers complementary to the adapter sequences (ligation-mediated PCR, LM-PCR). Exons were captured using the Agilent 38Mb SureSelect v2. In some cases, barcoded libraries from 2-4 subjects were mixed prior to hybridization with the capture reagent. After capture, another round of LM-PCR was performed to generate enough DNA to sequence. Libraries were sequenced using an IlluminaHiSeq2000.

At BCM, genomic DNA was sheared into fragments of approximately 120 base pairs with the Covaris S2 or E210 system. Fragments were processed through DNA End-Repair and A-tailing, and the resulting fragments were ligated with BCM-HGSC-designed Truncated-TA (TrTA) P1 and TA-P2 adapters with the NEB Quick Ligation Kit. Solid Phase Reversible Immobilization bead cleanup was used to purify the adapted fragments, after which nick translation and Ligation-Mediated PCR was performed using Platinum PCR Supermix HIFi. The

pre-capture libraries were hybridized to NimbleGen EZ Exome v2, VCRome v1, or VCRome v2.1 probes, either in solution or with solid-phase capture chips, and then amplied. In some cases, barcoded capture libraries were pooled in sets of 4 samples after postcapture amplication. Libraries were sequenced on the Life Technologies SOLiD platform using both 50bp fragment and 50x35bp paired-end run formats.

Data quality control and filtering

BI data was processed with Picard (http://picard.sourceforge.net/), which utilizes base quality score recalibration and local realignment at known indels and BWA for mapping reads to hg19. SNPs were called using GATK [40]. BCM data was processed with Picard and reads mapped to hg18 using Bfast [41]. The quality score recalibration and indel realignment was performed using GATK, followed by SNV identification using AtlasSNP 2 software [42]. Genotyping data from Affymetrix 5.0 and 6.0 was filtered using an MAF threshold of ≥5% and missing genotypes with ≤2% using PLINK and concordance checks were performed on the variant calls from the sequencing and genotyping arrays. 3 samples with low concordance between the exome sequencing and genotyping arrays (≤90%) were detected in the BI case-control dataset and discarded from further analyses.

The variants used in this study were restricted to sites that passed the standard GATK filters to eliminate SNPs with strand-bias, low quality for the depth of sequencing achieved, homopolymer runs, and SNPs near indels. And variants were required had an average read depth of ≥10x and a quality score of ≥30. Homozygous calls were required to have less than 10% of the alternate allele and heterozygous calls to have an allele balance of between 30% and 70%. A

HWE threshold of ≥0.05 was used as well. A set of 160 rare variants was selected for Sequenom validation and the validation rate using these filters was 99.5%.

Annotation and analyses

For the case-control datasets, we annotated each variant according to the longest transcript from the RefSeq database. The trio and quartet datasets were annotated using a custom pipeline that was built on top of the Variant Effect Predictor [43] to allow more stringent filtering of annotation artifacts from the 1000 Genomes Project [19]. The cases and controls in the BI dataset was compared separately from the cases and controls in the BCM dataset before combining the results, to ensure that differences in sequencing technologies and platforms did not affect the results. Variants on the autosomes were filtered using MAF≤5% in the controls from each dataset.

Variants on the X-chromosome were filtered using similar thresholds as the autosomal variants. In addition, variants that were found to be heterozygous in males were removed from the analyses as such inconsistencies were most likely to have resulted from mis-alignment errors. To increase the number of observations for the X-chromosome analyses, male probands from the trios/quartets were added as additional cases to the overall counts from the case-control datasets and their fathers were added as additional controls, since male offspring do not inherit their X-chromosomes from their fathers and the X-chromosomes in their fathers would serve as perfect normal controls. In addition, the MAF for rare variants on the X-chromosome were calculated from a large set of control females from the NHLBI exome sequencing study.

Fluidigm genotyping

96 PCR primer pairs and probes were designed by Fluidigm Corporation to amplify candidate mutations with a target amplicon size of 200 bp. Using the Fluidigm microfluidic platform, 96 multiplex-PCR reactions were performed using 96 DNA samples. Genotyping and clustering of calls were performed by manufacturer instructions for the Fluidigm Dynamic Array system in which assays are based on allele-specific PCR SNP detection chemistry and integrated fluidic circuits (IFCs).

Sanger sequencing

We designed primers and amplified regions around the candidate LoF sites according to standard protocols. PCR products were sequenced using traditional Sanger fluorescent dideoxy method on ABI 3730 capillary sequencers. Resulting sequences were analyzed and SNVs detected using SNPdetector software [44].

Sample preparation and pooling for Fluidigm PCR and MiSeq sequencing

The baseline concentration of genomic DNA was quantified by Quant-iT PicoGreen dsDNA reagent and detected on the Thermo Scientific Varioskan Flash. All DNAs were normalized to 50ng/μl and repeat quantification was performed to assess accuracy of the normalization step. The quantification and normalization was repeated again to ensure that all samples fell within the desired concentration range. A 10% variance was allowed, as that is the limit of quantitation of PicoGreen detection system. The normalization steps were done with robotic automation using the Packard Multiprobe II HT EX and Caliper LabChip GX system. Equimolar amounts of each DNA in a pool of samples is essential thus the same robotic

automation was used to guarantee a uniform pipetting error across all samples in all steps. Once each individual sample was normalized to 50ng/ul, 2 parental samples from different families were pooled together using a Multiprobe or Packard Robotic to total 25 pools (50 people). These pools along with individual probands were sent for Fluidigm PCR and MiSeq sequencing efforts described below.

Fluidigm PCR and MiSeq sequencing

Validation of selected variants was performed by targeted resequencing using microfluidic PCR (Access array system, Fluidigm) and the MiSeq sequencing system (Illumina). Father, mother and probands from each family with a LoF mutation were selected based on the presence of the indicated variant by whole exome sequencing. Target specific primers were designed to flank sites of interest and produce amplicons of 150-200 bp ± 20 bp. Molecularly barcoded, Illumina-compatible specific oligos, containing sequences complementary to the primer tails were added to the access array chip in the same well as the genomic DNA samples (20–50 ng of input) such that all amplicons for a given genomic sample share the same index. PCR was performed on the Fluidigm access array according to the manufacturer's instructions. Indexed libraries were recovered for each sample in a single collection well from the Fluidigm chip, quantified using PicoGreen, and then normalized for uniformity across libraries. Resulting normalized libraries were loaded on the MiSeq instrument and sequenced using paired end 150 bp sequencing reads. Paired-end sequencing was carried out by using MiSeq sequencing instruments; the resulting data were analyzed with the current Illumina pipeline. Standard quality control metrics, including error rates, percentage passing filter reads, and total Gb produced, were used to characterize process performance before downstream analysis. The Illumina

pipeline generates data files (BAM files) that contain the reads together with quality parameters. Detection of the presence of the targeted variants in each sample was done using the *mpileup* option in Samtools and were visually inspected using the *tview* option [45].

Linkage disequilibrium-based phasing of variant pairs

We adopted a linkage disequilibrium (LD) based method, similar to the four-haplotype test used to detect a recombination event, to phase pairs of variants within the same gene and applied this approach to predict compound heterozygous variants in the case-control datasets. A pair of variants (A and B) was predicted to occur on different chromosomes if:

1. We observed at least 1 individual who is heterozygous for variant A; and,
2. we observed at least 1 individual who is heterozygous for variant B; and,
3. we did not observe any individual who is homozygous at 1 variant and has at least 1 copy of the second variant (Figure 2.1).

In addition, since we cannot accurately phase singletons, we included all pairs of variants if at least one of them is a singleton.

Statistical analyses for global enrichment

For each variant, we calculated the MAF of the variant in the controls. The MAF of a variant pair is the maximum MAF of either variant in the pair. Multiple variant pairs within the same gene in the same individual were counted as a single complete knockout event. We calculated the normalized enrichment ratio as the (total number of events in cases/total number of events in controls)×(number of controls/number of cases) to handle the imbalance in the number of cases and controls that were sequenced. We assessed the statistical significance of the

global enrichment by shuffling the case-control labels for 10,000 permutations. For the enrichment analyses on the X-chromosome, one-sided hypergeometric probabilities were calculated assuming that hemizygous synonymous variants in male cases and controls are largely neutral variants. All the analyses were performed within each case-control dataset separately before combining the results, to ensure that the observations were not driven by a single dataset.

Copy number variant calling from exome sequencing data

The XHMM exome sequencing CNV discovery and genotyping pipeline [23] was run on these samples to detect exon-level copy number variation and assign CNV quality metrics. Stringent call-level QC was performed by removing all sex chromosome CNV and low-quality XHMM calls (XHMM SQ<60). This was followed by removal of outlier samples (those with no CNV, >50 CNV calls, or >5 MB of total CNV length).

Normalization of IQ scores from Raven's Colored Progressive Matrices (CPM)

To obtain a normalized IQ score, we performed linear regression on the CPM total scores for 151 cases between the ages of 4 to 11 who were part of this exome sequencing study, corrected for their ages ($\beta = 1.34$, $SE = 0.31$, $P = 3.48\times10^{-5}$).

ACKNOWLEDGEMENTS

We are most grateful to the families from all participating studies: Autism Genetic Resource Exchange (AGRE), the Autism Simplex Collection (TASC), National Database for Autism Research (NDAR), Boston Autism Consortium (AC) and Simons Simplex Collection (SSC). This work was directly supported by NIH grants R01MH089208 (MJD), R01 MH089025 (JDB), R01 MH089004 (GS), R01MH089175 (RG) and R01 MH089482 (JSS) and supported in part by NIH grants P50 HD055751 (EHC), RO1 MH057881 (BD), and R01 MH061009 (JSS). We thank Thomas Lehner (NIMH), Adam Felsenfeld (NHGRI), and Patrick Bender (NIMH) for their support and contribution to the project. EB, JDB, BD, MJD (communicating PI), RG, KR, AS, GS, JSS are lead investigators in the ARRA Autism Sequencing Consortium (ASC). We would also like to thank the NHLBI GO Exome Sequencing Project and its ongoing studies which produced and provided exome variant calls for comparison: the Lung GO Sequencing Project (HL-102923), the WHI Sequencing Project (HL-102924), the Broad GO Sequencing Project (HL-102925), the Seattle GO Sequencing Project (HL-102926) and the Heart GO Sequencing Project (HL-103010).

REFERENCES

1. Autism and Developmental Disabilities Monitoring Network Surveillance Year 2008 Principal Investigators CDC (2012) Prevalence of autism spectrum disorders--Autism and Developmental Disabilities Monitoring Network, 14 sites, United States, 2008. MMWR Surveill Summ 61: 1-19.
2. Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, et al. (2007) Strong association of de novo copy number mutations with autism. Science 316: 445-449.
3. Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, et al. (2008) Association between microdeletion and microduplication at 16p11.2 and autism. N Engl J Med 358: 667-675.
4. Sanders SJ, Ercan-Sencicek AG, Hus V, Luo R, Murtha MT, et al. (2011) Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. Neuron 70: 863-885.
5. Levy D, Ronemus M, Yamrom B, Lee YH, Leotta A, et al. (2011) Rare de novo and transmitted copy-number variation in autistic spectrum disorders. Neuron 70: 886-897.
6. Pinto D, Pagnamenta AT, Klei L, Anney R, Merico D, et al. (2010) Functional impact of global rare copy number variation in autism spectrum disorders. Nature 466: 368-372.
7. O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, et al. (2012) Sporadic autism exomes reveal a highly interconnected protein network of de novo mutations. Nature.
8. Iossifov I, Ronemus M, Levy D, Wang Z, Hakker I, et al. (2012) De novo gene disruptions in children on the autistic spectrum. Neuron 74: 285-299.
9. Neale BM, Kou Y, Liu L, Ma'ayan A, Samocha KE, et al. (2012) Patterns and rates of exonic de novo mutations in autism spectrum disorders. Nature.

10. Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, et al. (2012) De novo mutations revealed by whole-exome sequencing are strongly associated with autism. Nature.

11. Constantino JN, Todorov A, Hilton C, Law P, Zhang Y, et al. (2012) Autism recurrence in half siblings: strong support for genetic mechanisms of transmission in ASD. Mol Psychiatry.

12. Ritvo ER, Spence MA, Freeman BJ, Mason-Brothers A, Mo A, et al. (1985) Evidence for autosomal recessive inheritance in 46 families with multiple incidences of autism. Am J Psychiatry 142: 187-192.

13. Jorde LB, Hasstedt SJ, Ritvo ER, Mason-Brothers A, Freeman BJ, et al. (1991) Complex segregation analysis of autism. Am J Hum Genet 49: 932-938.

14. Zweier C, de Jong EK, Zweier M, Orrico A, Ousager LB, et al. (2009) CNTNAP2 and NRXN1 are mutated in autosomal-recessive Pitt-Hopkins-like mental retardation and determine the level of a common synaptic protein in Drosophila. Am J Hum Genet 85: 655-666.

15. Morrow EM, Yoo SY, Flavell SW, Kim TK, Lin Y, et al. (2008) Identifying autism loci and genes by tracing recent shared ancestry. Science 321: 218-223.

16. Casey JP, Magalhaes T, Conroy JM, Regan R, Shah N, et al. (2011) A novel approach of homozygous haplotype sharing identifies candidate genes in autism spectrum disorder. Hum Genet.

17. Chahrour MH, Yu TW, Lim ET, Ataman B, Coulter ME, et al. (2012) Whole-exome sequencing and homozygosity analysis implicate depolarization-regulated neuronal genes in autism. PLoS Genet 8: e1002635.

18. Girirajan S, Rosenfeld JA, Coe BP, Parikh S, Friedman N, et al. (2012) Phenotypic Heterogeneity of Genomic Disorders and Rare Copy-Number Variants. N Engl J Med.

19. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, et al. (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335: 823-828.

20. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI (2008) Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet 82: 100-112.

21. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.

22. Kang HJ, Kawasawa YI, Cheng F, Zhu Y, Xu X, et al. (2011) Spatio-temporal transcriptome of the human brain. Nature 478: 483-489.

23. Fromer M, Moran JL, Chambert K, Banks E, Bergen SE, et al. (2012) Discovery and Statistical Genotyping of Copy-Number Variation from Whole-Exome Sequencing Depth. Am J Hum Genet 91: 597-607.

24. Devlin B, Scherer SW (2012) Genetic architecture in autism spectrum disorder. Curr Opin Genet Dev.

25. Gottipati S, Arbiza L, Siepel A, Clark AG, Keinan A (2011) Analyses of X-linked and autosomal genetic variation in population-scale whole genome sequencing. Nat Genet 43: 741-743.

26. Zhang B, Kirov S, Snoddy J (2005) WebGestalt: an integrated system for exploring gene sets in various biological contexts. Nucleic Acids Res 33: W741-748.

27. Newschaffer CJ, Croen LA, Daniels J, Giarelli E, Grether JK, et al. (2007) The epidemiology of autism spectrum disorders. Annu Rev Public Health 28: 235-258.

28. Yan D, Liu XZ (2010) Genetics and pathological mechanisms of Usher syndrome. J Hum Genet 55: 327-335.

29. Johansson M, Gillberg C, Rastam M (2010) Autism spectrum conditions in individuals with Mobius sequence, CHARGE syndrome and oculo-auriculo-vertebral spectrum: diagnostic aspects. Res Dev Disabil 31: 9-24.

30. Celestino-Soper PB, Shaw CA, Sanders SJ, Li J, Murtha MT, et al. (2011) Use of array CGH to detect exonic copy number variants throughout the genome in autism families detects a novel deletion in TMLHE. Hum Mol Genet 20: 4360-4370.

31. Celestino-Soper PB, Violante S, Crawford EL, Luo R, Lionel AC, et al. (2012) A common X-linked inborn error of carnitine biosynthesis may be a risk factor for nondysmorphic autism. Proc Natl Acad Sci U S A 109: 7974-7981.

32. Nava C, Lamari F, Heron D, Mignot C, Rastetter A, et al. (2012) Analysis of the chromosome X exome in patients with autism spectrum disorders identified novel candidate genes, including TMLHE. Transl Psychiatry 2: e179.

33. Speevak MD, Farrell SA (2011) Non-syndromic language delay in a child with disruption in the Protocadherin11X/Y gene pair. Am J Med Genet B Neuropsychiatr Genet 156B: 484-489.

34. Ishikawa T, Miyata S, Koyama Y, Yoshikawa K, Hattori T, et al. (2012) Transient expression of Xpn, an XLMR protein related to neurite extension, during brain development and participation in neurite outgrowth. Neuroscience.

35. Roll P, Rudolf G, Pereira S, Royer B, Scheffer IE, et al. (2006) SRPX2 mutations in disorders of language cortex and cognition. Hum Mol Genet 15: 1195-1207.

36. Roll P, Vernes SC, Bruneau N, Cillario J, Ponsole-Lenfant M, et al. (2010) Molecular networks implicated in speech-related disorders: FOXP2 regulates the SRPX2/uPAR complex. Hum Mol Genet 19: 4848-4860.

37. Renieri A, Mari F, Mencarelli MA, Scala E, Ariani F, et al. (2009) Diagnostic criteria for the Zappella variant of Rett syndrome (the preserved speech variant). Brain Dev 31: 208-216.

38. Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, et al. (2008) Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320: 539-543.

39. Yu TW, Chahrour MH, Coulter ME, Jiralerspong S, Okamura-Ikeda K, et al. (2013) Using whole exome sequencing to identify inherited causes of autism. Neuron.

40. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, et al. (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20: 1297-1303.

41. Homer N, Merriman B, Nelson SF (2009) BFAST: an alignment tool for large scale genome resequencing. PLoS One 4: e7767.

42. Challis D, Yu J, Evani US, Jackson AR, Paithankar S, et al. (2012) An integrative variant analysis suite for whole exome next-generation sequencing data. BMC Bioinformatics 13: 8.

43. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, et al. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26: 2069-2070.

44. Zhang J, Wheeler DA, Yakub I, Wei S, Sood R, et al. (2005) SNPdetector: a software tool for sensitive and accurate SNP detection. PLoS Comput Biol 1: e53.

45. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, et al. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078-2079.

CHAPTER 3

A Population-based Approach for Detecting Rare Recessives Implicates the Cholesterol Biosynthesis gene *DHCR24* in Autism Spectrum Disorder and Intellectual Disability

A Population-based Approach for Detecting Rare Recessives Implicates the Cholesterol Biosynthesis gene *DHCR24* in Autism Spectrum Disorder and Intellectual Disability

Elaine T. Lim[1,2,3,4,¶], Yingleong Chan[2,3,4], Susanne Goetze[5], Daniel Spatt[3], Lisa Kratz[6], Phil H. Lee[1,2], Maria Chahrour[5], Matthew Johnson[5], Douglas M. Ruderfer[7], Jacqueline I. Goldstein[1,2], Christine Stevens[1,2], Shaun M. Purcell[7], Joel N. Hirschhorn[2,3,4], Soumya Raychaudhuri[8], Christopher A. Walsh[5], Federick Winston[3], Richard Kelley[6], Mark J. Daly[1,2,3,*], and Timothy W. Yu[5,*,¶]

[1] Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA 02114, USA.

[2] Program in Medical and Population Genetics, Broad Institute, Cambridge, MA 02142, USA.

[3] Departments of Genetics and Medicine, Harvard Medical School, Boston, MA 02115, USA.

[4] Program in Genetics and Genomics, Biological and Biomedical Sciences, Harvard Medical School, Boston, MA 02115, USA.

[5] Division of Genetics, Boston's Children Hospital, Boston, MA 02115, USA.

[6] Clinical Mass Spectrometry Lab, Kennedy Krieger Institute, Baltimore, MD 21205, USA.

[7] Department of Psychiatry, Mount Sinai School of Medicine, New York, NY 10029, USA.

[8] Division of Immunology, Allergy, and Rheumatology, Brigham and Women's Hospital, Boston, MA 02115, USA.

[*] These authors contributed equally.

[¶] Correspondence: elimtt@gmail.com, timothy.yu@childrens.harvard.edu

Manuscript in preparation.

ABSTRACT

We describe the development of a novel population-based methodology (named RAFT for **R**ecessive **A**llele **F**requency-based **T**est) for discovering rare recessive variants and genes in complex diseases using whole-exome sequencing or genotyping technologies. We demonstrate that our approach is better powered than conventional population-based recessive tests. In applying our methodology to a set of exome chip and exome sequencing studies, we discovered evidence for rare recessive missense variants in the cholesterol synthesis gene 3-β-Hydroxysteroid-Δ 24 Reductase (*DHCR24*) in autism and intellectual disability. We further performed a series of site-directed mutagenesis and desmosterol-to-cholesterol conversion assays to evaluate and characterize the functionality of these missense variants in cholesterol synthesis.

INTRODUCTION

Identifying the genetic causes of rare recessive Mendelian diseases using whole-exome sequencing has proven extremely successful [1]. Recent reports have also suggested a significant role for recessive alleles in the genetic architecture of complex diseases such as schizophrenia and autism spectrum disorders or ASDs [2,3,4,5]. However, identifying such rare recessive subtypes in complex diseases has been challenging, given the genetic heterogeneity and polygenicity of complex diseases [3]. Demonstrated approaches for screening rare recessive alleles in complex diseases include the use of unique populations such as consanguineous populations [4,6] or founder populations such as the Icelandic, Ashkenazi Jewish and Finnish populations [7,8] by identifying rare homozygous variants in regions with long runs of homozygosity [2,9]. Evaluation of the significance of the homozygous variants discovered from these screens has typically relied upon linkage calculations, or population-based approaches such

as chi-squared test, logistic regression or Fisher's Exact Test. Here, we utilize a population-based approach (which we named RAFT for "**R**ecessive **A**llele **F**requency-based **T**est") to prioritize homozygous variants in a common disease based on their rarity of these variants in terms of their allele frequencies (Materials and Methods). We have also adapted this population-based approach to family-based datasets to increase the power for discovering rare recessive genes in ASD. In applying our approach, we discovered a cholesterol biosynthesis gene (*DHCR24*) where milder missense variants in the gene contribute to ASD in a recessive mode of inheritance while more severe missense variants contribute to intellectual disability (ID), as well as brain malformations and fetal death.

RESULTS

Evaluating power for existing tests on simulated monogenic and polygenic diseases

We first considered a monogenic recessive disease where there is a single causal gene involved, for instance cystic fibrosis transmembrane conductance regulator (*CFTR*) in cystic fibrosis (CF). If we were to sequence 40 individuals with CF and observed that half of them were homozygous for the ΔF508 deletion in *CFTR*, and that none in 4,000 unaffected controls were homozygous for ΔF508, the p-value obtained for this observation using Fisher's Exact Test on the homozygous counts (Hom-FET) is highly significant ($<1\times10^{-15}$). Next, we consider an oligogenic recessive disease where there are several causal genes involved, for instance Usher Syndrome where there are 20 causal genes known to-date. Now if we were to sequence 40 cases we might only observe 2 cases who are recessive (either homozygous or compound heterozygous) for LoF variants in one of the genes, for instance Usher Syndrome 2A (*USH2A*). The p-value obtained from Fisher's Exact Test is far more modest (9.6×10^{-5}) for this observation

(Figure 3.1). Given the polygenicity involved in a complex disease such as ASD where we expect hundreds of genes and several modes of inheritance to be involved, we would have very little power to detect individual recessive genes contributing to a small percentage of cases, even if they were fully penetrant.

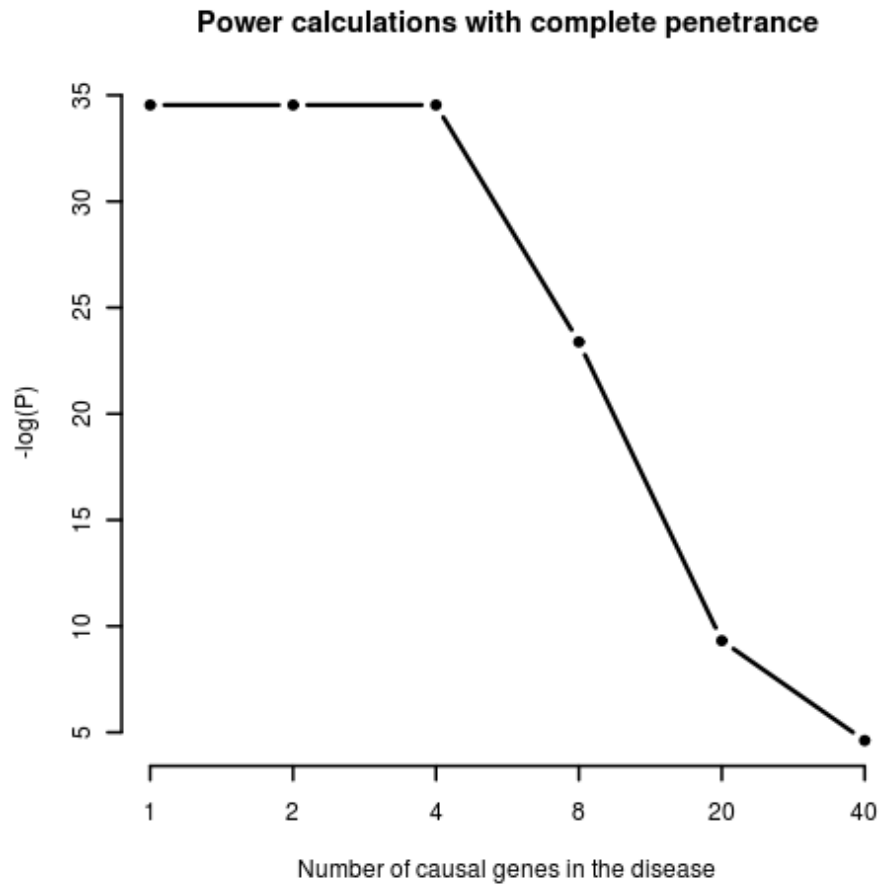
To further illustrate the effect of polygenicity on recessive gene discovery, we used a previously published instance where 2 independent families of Middle Eastern ancestry with ASD, intellectual disability and epilepsy were reported to harbor rare LoF variants in the *BCKDK* gene [10]. Given the background where ~200 families of Middle Eastern ancestry with ASD have been sequenced to-date [4], the p-value obtained using Fisher's Exact Test for observing 2 independent probands with recessive LoF variants in *BCKDK* out of 200 ASD cases of Middle Eastern ancestry, compared to none with recessive variants in *BCKDK* among the 400 parents, is only 0.11 – an unremarkable result in an exome-wide screen of all genes. However, we demonstrate later on that we would have identified *BCKDK* from an exome-wide screen for recessive genes using our population-based approach.

Figure 3.1: Power Calculations to Illustrate Decreased Power in Complex Diseases

(A) P-values obtained when performing Fisher's Exact Test on the homozygous counts (Hom-FET) if there are 1, 2, 4, 8, 20 or 40 genes with complete penetrance that are causal for the disease. (B) P-values obtained using Hom-FET with different percentages of variance explained by the homozygotes (aa) in the cases, assuming complete penetrance. This is an ideal scenario and the power for detection will be lower if the variants exhibit incomplete penetrance or if there are fewer controls available.

Figure 3.1: Power Calculations to Illustrate Decreased Power in Complex Diseases (Continued)

A



B

Example	% variance explained	E(Case aa)	E(Case Aa or AA)	E(Control aa)	E(Control Aa or AA)	Hom-FET E(p-value)
CFTR in cystic fibrosis	50%	20	20	0	4,000	$<1\times10^{-15}$
USH2A in Usher Syndrome	5%	2	38	0	4,000	9.6×10^{-5}
Gene X in ASD	2.5%	1	39	0	4,000	1.0×10^{-2}

RAFT test statistics

It has been proposed previously that incorporating the departure from Hardy-Weinberg equilibrium (HWE) can improve the power to detect such rare recessives [11,12,13]. As such, we developed a likelihood-based association test (RAFT) to evaluate jointly the significance of the excess homozygosity from the expected under HWE, as well as the significance of the case-control deviation. Using conventional case-control recessive tests, an observation of 5 cases and 0 controls being homozygous for a variant of 0.5% allele frequency is no more significant than the same 5 to 0 observation for a variant of 5% allele frequency. However, the intuition behind our test is to assess how unusual is an observation with *N* number of recessives in the cases given the expected number of recessives based on the allele frequency. As such, we would expect the observation of 5 cases being homozygous for a variant of 0.5% allele frequency to be far more significant than 5 cases being homozygous for a variant of 5% allele frequency.

However, for particularly rare variants, it is expected that many recessive cases are compound heterozygotes rather than homozygous in nature. Following our previous analysis [3], we modified the RAFT statistic using the composite allele frequency of variants in a target class within the same gene, for instance, all LoF variants, or all non-synonymous (missense and LoF) variants. We used this composite allele frequency to estimate the probability that a single chromosome drawn from a population carries an alternate allele in the gene and this can be used to estimate the probability under a recessive mode of inheritance (homozygous and compound heterozygous) using the modified RAFT statistic. While this will very likely be an effective screening strategy for LoF variants, it is possible that for rare missense variants, the inevitable inclusion of neutral variants will reduce the power [14].

Using the previous example with the 2 recessive LoF observations in *BCKDK*, the composite allele frequency (or sum of the allele frequencies) of all LoF variants in *BCKDK* is 0%, i.e. there are no LoF variants found in 207 unrelated Middle Eastern individuals. The p-value calculated using RAFT for these 2 recessive observations in *BCKDK* is 2.0×10^{-9} (Supplementary Methods), which is exome-wide significant at an α threshold of 0.05 after Bonferroni correction for 17,974 genes (P-value threshold = 2.8×10^{-6}). Moreover, this observation is even more significant after including each additional affected sibling in those 2 families who are homozygous for the LoF variants as replication (P×0.25 for each additional sibling).

Application of RAFT to autism datasets

We first applied our approach to a whole-exome genotyping dataset comprising of 1,069 unrelated individuals of European ancestry with ASD as "cases" and another 2,141 unrelated parents of matched ancestry without ASD as "controls" from the Autism Genetic Resource Exchange (AGRE) collection by first ensuring that the cases and controls were from a homogeneous population of European ancestry (Figure 3.2). In the process, we discovered a rare missense variant (P244L) that is found at 0.05% allele frequency (Table 3.1, Figure 3.2) segregating perfectly in an autosomal recessive mode of inheritance within a single family of 3 affected children (OR>4000, $P = 7.8\times10^{-6}$, Figure 3.3A). Given the allele frequency for the missense variant, we expected to observe a single homozygote in 4 million individuals.

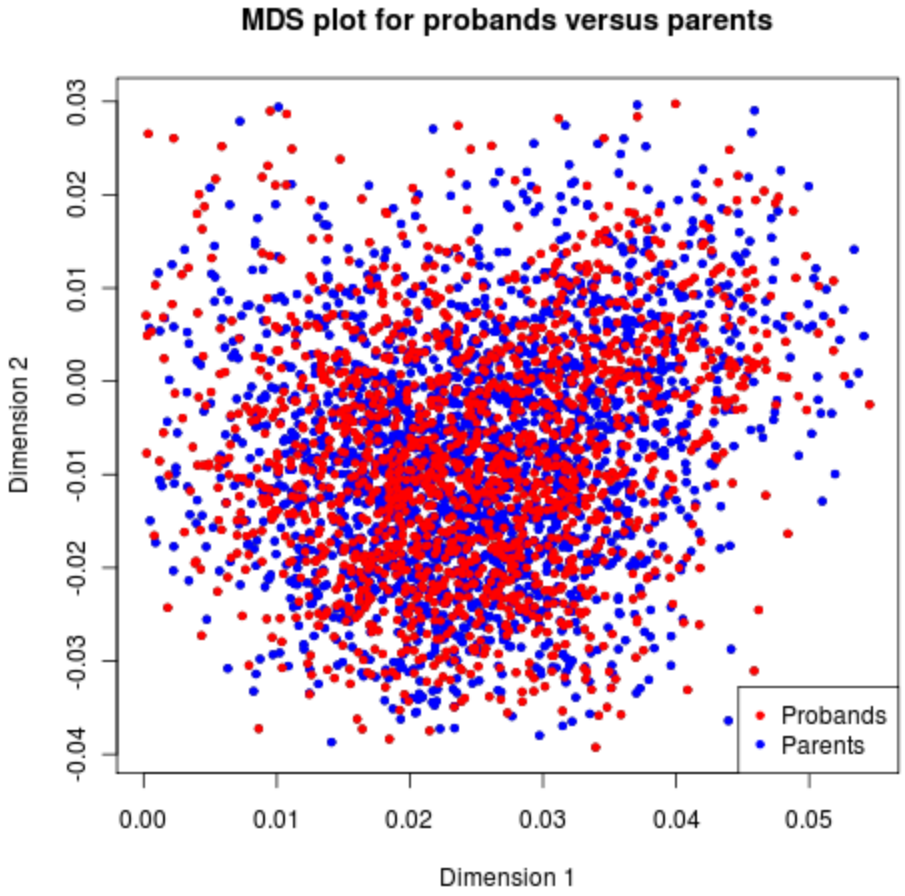


Figure 3.2: The first 2 dimensions using MDS for the individuals in the exome chip study

The MDS plot for the probands (one unrelated affected child from each family) and both parents from each family of European ancestry from the exome chip genotyping data.

Table 3.1: Top hits from the exome chip scan

The odds ratio, LOD score and P-values calculated using RAFT, as well as the p-value calculated using Fisher's Exact Test on the allele counts (FET P-value) and Fisher's Exact Test on the homozygous counts (Hom-FET P-value) shown. For the *DHCR24* P244L variant, there are another 2 affected children in the same family who are homozygous for the same rare variant, so the final p-value was multiplied by another $0.25 \times 0.25 = 7.8 \times 10^{-6}$.

Chr	Position	Case AA	Case Aa	Case aa	Control AA	Control Aa	Control aa	Case Allele Freq	Control Allele Freq	Odds Ratio	LOD score	P-value	Gene	Variant	Amino Acid	FET P-value	Hom-FET P-value
16	17228363	1067	1	1	2140	1	0	0.0014	0.0002	9636	3.55	0.00005	XYLT1	missense	p.T665M	0.11	0.33
1	32087170	1067	1	1	2139	2	0	0.0014	0.0005	4281	3.20	0.00013	HCRTR1	missense	p.F239L	0.34	0.33
1	55337168	1067	1	1	2137	2	0	0.0014	0.0005	4276	3.20	0.00013	DHCR24	missense	p.P244L	0.34	0.33
22	41077895	1057	2	1	2121	1	0	0.0019	0.0002	4243	3.19	0.00013	MCHR1	missense	p.T411M	0.05	0.33
11	69490001	1066	2	1	2139	2	0	0.0019	0.0005	2408	2.95	0.00023	ORAOV1	missense	p.G3S	0.1	0.33
12	14577265	1067	1	1	2138	3	0	0.0014	0.0007	2408	2.95	0.00023	ATF7IP	missense	p.L139R	0.41	0.33
15	72030269	1067	1	1	2138	3	0	0.0014	0.0007	2408	2.95	0.00023	THSD4	missense	p.P610L	0.41	0.33
2	242312656	1065	1	1	2133	3	0	0.0014	0.0007	2402	2.95	0.00023	FARP2	missense	p.H45R	0.41	0.33
11	47311032	1067	1	1	2137	4	0	0.0014	0.0009	1540	2.75	0.00037	MADD	missense	p.S892C	0.69	0.33
15	78572759	1066	2	1	2138	3	0	0.0019	0.0007	1540	2.75	0.00037	WDR61	downstream	NA	0.23	0.33
16	49671518	1066	2	1	2138	3	0	0.0019	0.0007	1540	2.75	0.00037	ZNF423	silent	p.N515N	0.23	0.33
21	34897281	1067	1	1	2137	4	0	0.0014	0.0009	1540	2.75	0.00037	GART	missense	p.L365V	0.69	0.33
2	97279225	1067	1	1	2137	4	0	0.0014	0.0009	1540	2.75	0.00037	KANSL3	silent	p.P265P	0.69	0.33
2	170366485	1067	1	1	2137	4	0	0.0014	0.0009	1540	2.75	0.00037	KBTBD10	missense	p.I66T	0.69	0.33
4	155254512	1067	1	1	2137	4	0	0.0014	0.0009	1540	2.75	0.00037	DCHS2	missense	p.V451L	0.69	0.33
2	67630980	550	406	113	1115	859	167	0.3	0.28	1.4	2.65	0.00048	ETAA1	missense	p.S389N	0.16	0.01
1	89123443	414	461	194	805	1019	317	0.4	0.39	1.3	2.64	0.00048	-	-	-	0.4	0.02
11	45949745	1067	1	1	2136	5	0	0.0014	0.0012	1069	2.60	0.00055	GYLTL1B	missense	p.R591Q	0.73	0.33
11	102076653	1067	1	1	2136	5	0	0.0014	0.0012	1069	2.60	0.00055	YAP1	missense	p.P278S	0.73	0.33

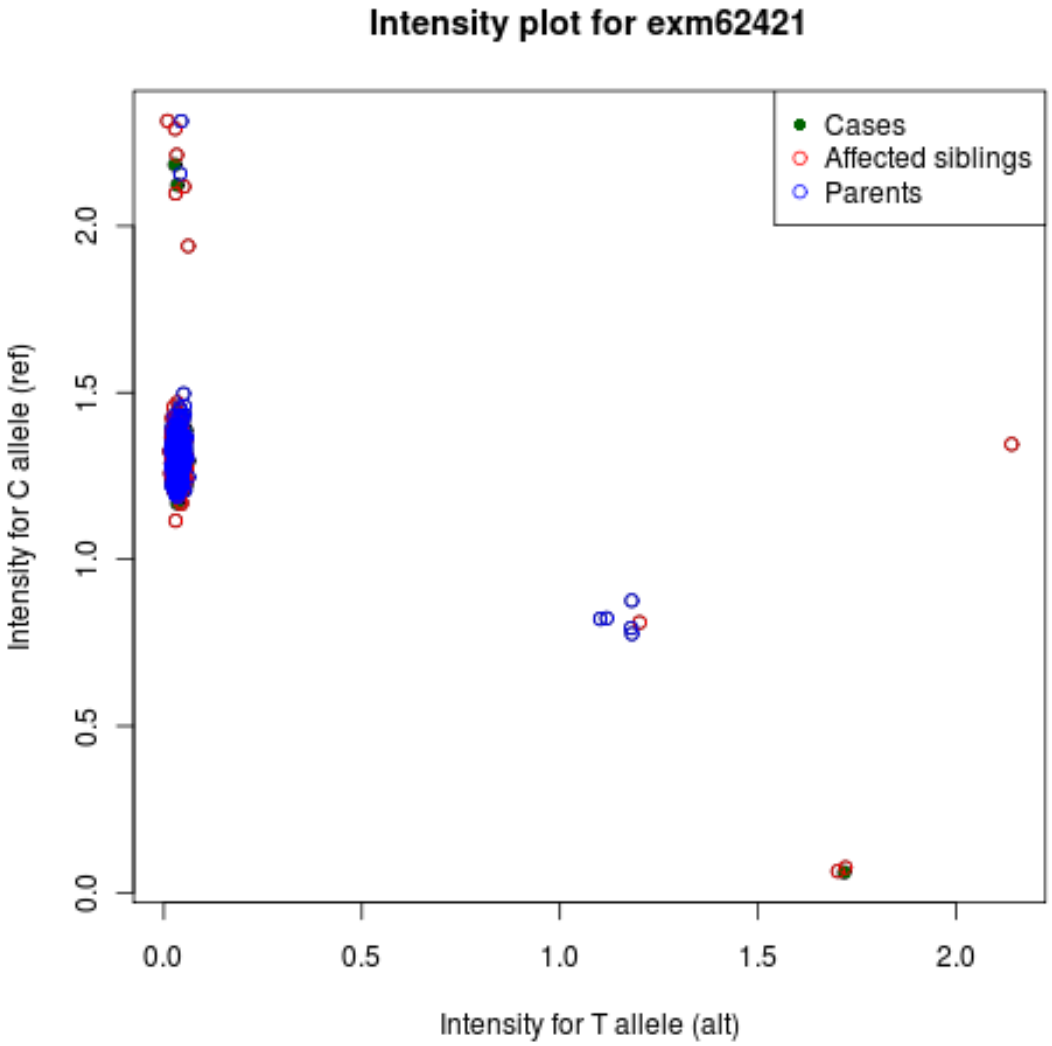


Figure 3.2: Intensity plot for the *DHCR24* P244L variant discovered in the Boston family

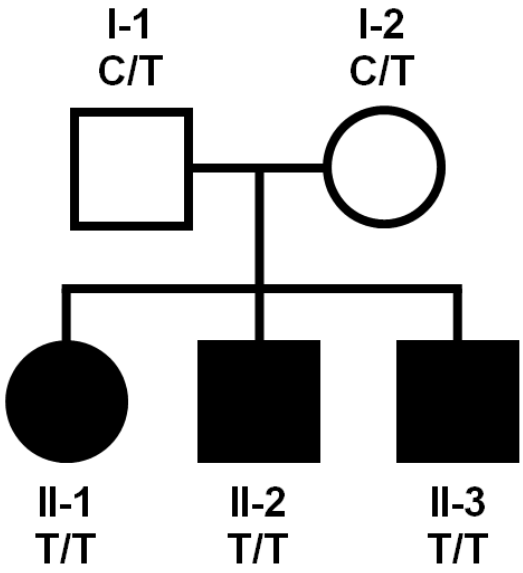
The intensities for the probands (unrelated cases) are colored in green, affected siblings colored in red and parents in blue.

Figure 3.3: Pedigrees for the 3 families with autism and intellectual disability and homozygous variants in *DHCR24*

(A) Pedigree for the first European family with ASD, (B) pedigree for the second Middle Eastern family with ASD and (C) pedigree for the third Middle Eastern family with intellectual disability.

Figure 3.3: Pedigrees for the 3 families with autism and intellectual disability and homozygous variants in *DHCR24* (Continued)

A



B

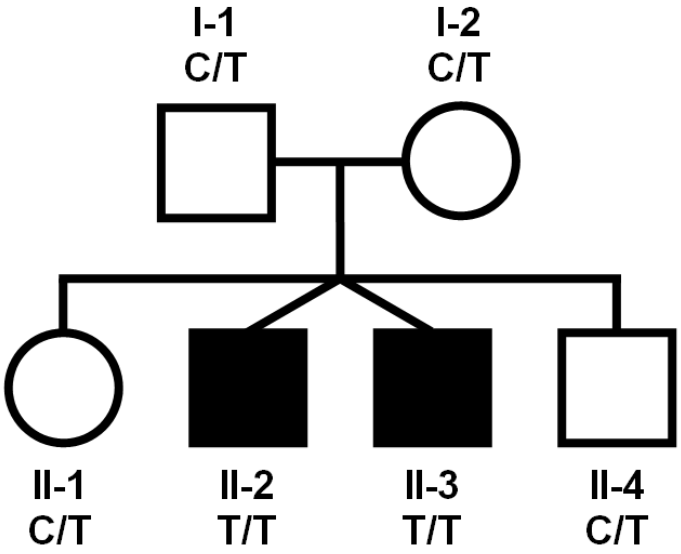
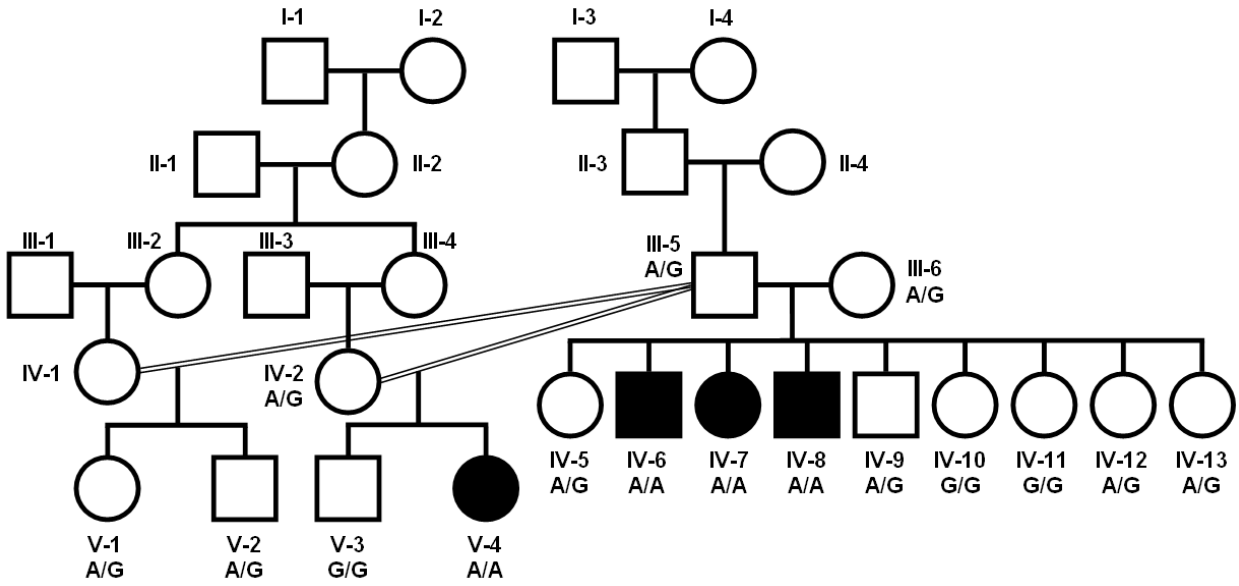


Figure 3.3: Pedigrees for the 3 families with autism and intellectual disability and homozygous variants in *DHCR24* (Continued)

C



This prompted us to scan additional exome sequencing datasets, including 207 Middle Eastern families with a broad range of neuro-developmental disorders including autism and intellectual disability. We discovered two additional families, both of Middle Eastern ancestry, with rare homozygous missense variants in *DHCR24* that segregated perfectly in a recessive mode of inheritance within the families. Given that the composite allele frequency of LoF and missense variants in *DHCR24* in European Americans is 0.44% and 0.48% in Middle Eastern individuals, the RAFT p-value obtained for these 3 recessive observations is 7.64×10^{-10} (Supplementary Methods), which is exome-wide significant.

The second family was a Pakistani family with a pair of monozygotic twins diagnosed with autism (recruited as part of the AGRE collection), both of whom were homozygous for a R478Q missense change (Figure 3.3B). The third family was a large consanguineous pedigree from Saudi Arabia and had with four children diagnosed with intellectual disability (Figure 3.3C). The affected children in this family were homozygous for a R103H missense change. Both the R478Q and R103H variants were not found in 400 independent individuals of Middle Eastern ancestry.

The three missense variants discovered in these three families with ASD and intellectual disability (P244L, R478Q and R103H) all affect amino acid residues that are highly conserved in 46 vertebrates, and all are predicted by PolyPhen2 to be “probably damaging”. A lookup in 6,500 control individuals from the Exome Variant Server, as well as 1,000 individuals of Middle Eastern ancestry with exome sequencing data and an exome sequencing dataset comprising of 26,000 individuals revealed no other individuals homozygous for these variants, nor any other rare (<1% allele frequency), protein-altering (missense, nonsense, splice site, or frameshift) variants in *DHCR24* (Fisher’s Exact Test $P = 1.8\times10^{-4}$).

The 24-dehydrocholesterol reductase gene (*DHCR24*), a gene involved in cholesterol synthesis by converting desmosterol into cholesterol. Rare recessive missense variants in *DHCR24* had been previously reported in 4 independent families affected by a rare disorder called desmosterolosis, which resulted in a variety of recognizable phenotypic manifestations such as microcephaly, agenesis or thinning of the corpus callosum and polydactyly [15,16,17,18]. Overlapping physical manifestations are seen in other disorders of cholesterol biosynthesis such as Smith-Lemli-Optiz Syndrome (SLOS), but the key distinguishing characteristic seen in the affected individuals was unusually high levels of desmosterol compared to total cholesterol in their serum. In one of these families, recessive missense variants in *DHCR24* resulted in infant death shortly after birth [18], and by analogy to SLOS, many severe LoF variants in the gene are likely to be inviable in humans. Post-mortem analysis of the percentage of accumulated desmosterol in this case was measured across three different tissues (liver, kidney and brain) and found to be the highest in the brain, reaffirming the importance of cholesterol in the developing brain.

In addition, we discovered a family with two fetal demises and both fetuses harbored compound heterozygous variants in *DHCR24* (C36X and P443L). This observation confirmed that complete loss of *DHCR24* (such as from the C36X variant) confers recessive lethality in humans and in some instances, severe missense variants such as the P443L variant, as well as the N294T, K306N and Y471S variants that were previously reported, can similarly result in recessive lethality. All 5 variants were not found in 33,500 control individuals and are likely to be private to the families studied. Interestingly, we found a rare protein-truncating variant (Y237X) that is found at 0.03% in European Americans from the Exome Variant Server and estimate that at least 1 in 1,700 individuals harbor a LoF variant in *DHCR24* and this translates

to a recessive lethality rate of at least 1 in 11 million from a complete loss of *DHCR24*. Notably, the missense variant discovered in these two fetal demises (P443L) is in close proximity with the 3 previously reported missense variants that resulted in fetal death and these variants are located near the p53 binding domain in *DHCR24*, and that region in the protein has been proposed to be involved in binding to desmosterol [19], suggesting that there might be a yet undiscovered domain in that region of the gene that is critical for human viability.

Functional evidence for *DHCR24* variants

Since *DHCR24* is likely to influence disease manifestations in a recessive manner, we performed an allelic screen of all the 16 missense variants and 1 LoF variant found in *DHCR24* from the European Americans in the Exome Variant Server (Figure 3.4), in order to assess the distribution of cholesterol synthesis for the variants, as well as to estimate the carrier rate of European Americans harboring a LoF variant or a missense variant predicted to be deleterious from the allelic screen. We performed the allelic screen by creating mutants of all the 17 variants, as well as the 2 additional variants from the families with ASD and intellectual disability, the 2 variants from the fetal demises and 3 reported variants from previous desmosterolosis cases, resulting in a total of 24 variants tested. As controls, we included yeast transformed with wildtype *DHCR24* and yeast transformed with the plasmid alone. We have performed site-directed mutagenesis for all mutants and are currently inducing protein expression for all the *DHCR24* mutants and optimizing a desmosterol-to-cholesterol biochemical assay adapted from Waterham *et al.* [20], in order to evaluate and assess the functionality of the variants discovered in the cases with autism and intellectual disability, as well as a population-based survey of missense variants in the gene.

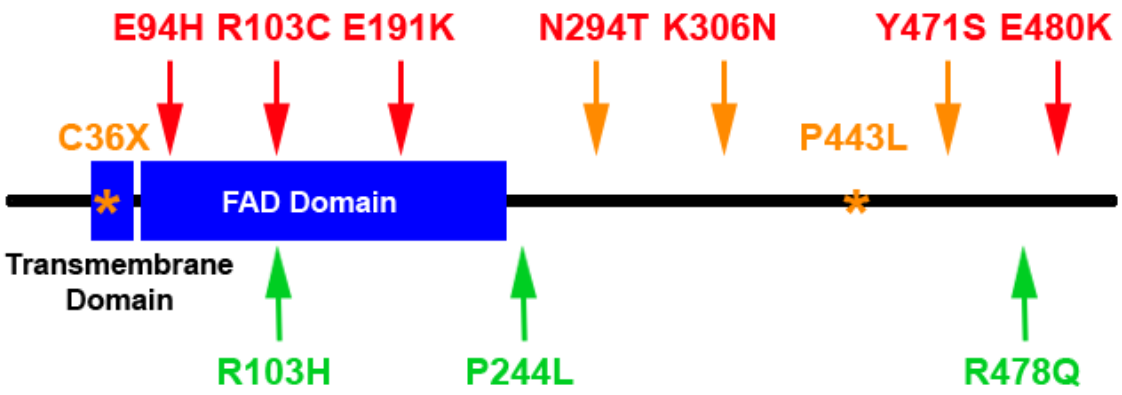
Figure 3.4: Variants in *DHCR24* discovered to-date

(A) The previously reported disease-causing variants in patients with desmosterolosis highlighted in red and shown on top; the variants discovered in the individuals with autism and intellectual disability are highlighted in green; and the variants discovered to result in human lethality highlighted in orange or marked by orange arrows.

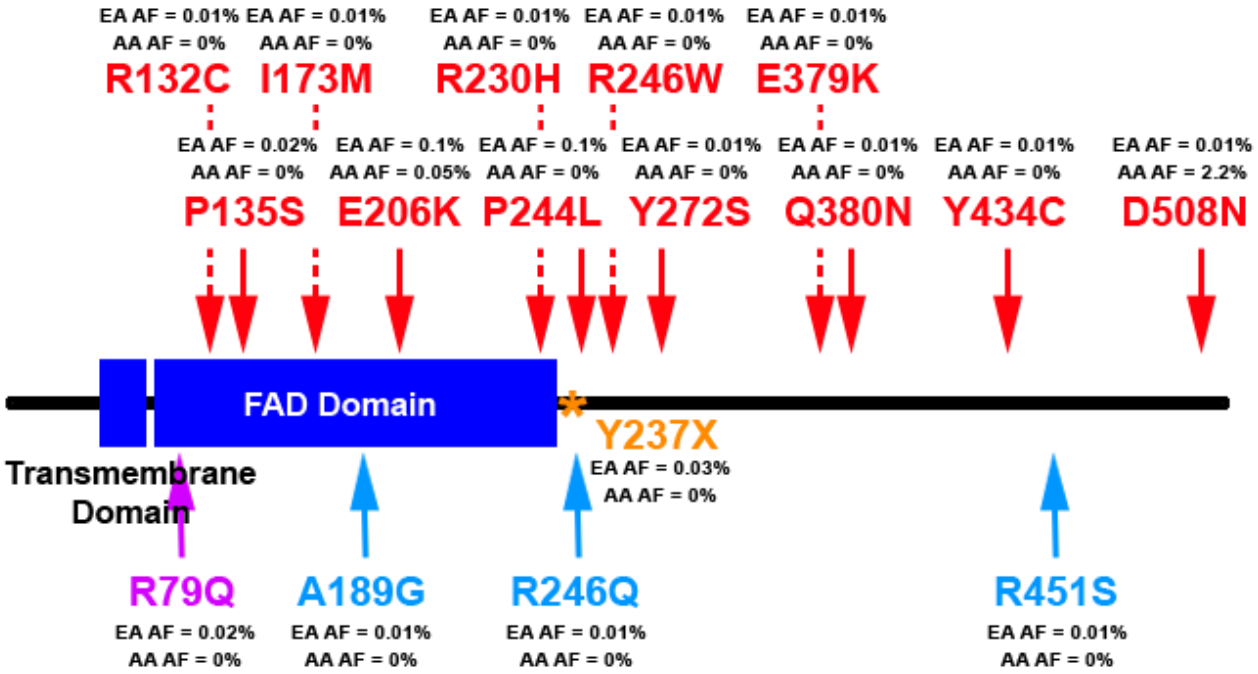
(B) Diagram of the variants found in the 4,300 European Americans from the Exome Variant Server with the allele frequencies of the variants shown for European Americans (EA AF) and African Americans (AA AF). The variants colored in red are predicted by PolyPhen2 to be "probably damaging", the variants colored in purple are predicted to be "possibly damaging", the variants colored in blue are predicted to be "benign" and loss-of-function (LoF) variants are colored in orange.

Figure 3.4: Variants in *DHCR24* discovered to-date (Continued)

A



B



DISCUSSION

Cholesterol is known to be important for the formation of myelin sheaths and membrane lipid rafts, and is a signaling molecule involved in developmental pathways such as the Sonic Hedgehog signaling pathway that regulates limb development. Recessive mutations in several genes that affect the precursors for cholesterol synthesis have been implicated in various neuro-developmental diseases, such as Smith–Lemli–Opitz syndrome, desmosterolosis and lathosterolosis [21]. *DHCR24* (also known as Selective Alzheimer's Disease Indicator-1 or *Seladin-1*) is gene that encodes a 516-amino acid long protein and is found on 1p32.3, where 2 rare duplications and one *de novo* deletion were previously reported in individuals with ASD [22]. The gene was also named *Seladin-1* based on the initial discoveries that *DHCR24* was down-regulated in parts of the brain known to be important in Alzheimer's Disease [23,24], although this association has been debated recently [25,26]. The proportion of individuals with ASD and low cholesterol levels has been estimated to be as high as 20% [27] and a recent study in an ASD-related disorder (Rett Syndrome) has also highlighted the importance of the cholesterol synthesis pathway in neuro-developmental disorders [28].

In this study, we present evidence for rare recessive variants in a critical gene involved in cholesterol synthesis (*DHCR24*) contributing to ASD and intellectual disability, as well as provide additional evidence that complete loss or severe missense variants in *DHCR24* can result in recessive lethality in humans. In addition to describing a novel statistical screen (RAFT) to identify such rare recessive variants in a complex disorder such as ASD, we are evaluating functional evidence for the pathogenicity of the 3 new variants discovered in families with ASD and intellectual disability (R103H, P244L and R478Q) using a cholesterol synthesis assay in yeast. In addition, we also surveyed the missense variants found in a large control population of

4,300 Europeans and estimate that approximately 1 in 150 individuals are carriers for a pathogenic variant in *DHCR24* and that approximately 1 in 86,000 individuals carry recessive pathogenic variants in *DHCR24*, although the rate of recessive carriers is much higher in ASD and intellectual disability (given that we discovered 3 independent families with recessive variants in *DHCR24* among approximately 3,000 families).

We discovered that knocking out *DHCR24* has severe consequences on fetal lethality and drugs that target the gene are likely to result in serious side effects. However, cholesterol supplements, such as egg yolks, have been shown to aid improvements in children with epilepsy and other brain disorders such as Smith-Lemli-Opitz syndrome, and such dietary treatment can potentially help these individuals with ASD and intellectual disability as well. There are likely to be other genes that are directly involved in the cholesterol synthesis pathway, or are involved in regulating and modulating the cholesterol synthesis pathway, that can similarly confer a significant risk to ASD and intellectual disability. The identification and understanding of such genes, as well as the identification of individuals affected by cholesterol deficiency or accumulation of the precursor substrates, can help provide immediate dietary treatment for these individuals, as well as further our understanding of the different components involved in regulating cholesterol synthesis for therapeutic development.

MATERIALS AND METHODS

Exome chip quality control and analyses

Genotyping was conducted using the Illumina HumanExome bead chip at the Broad Institute on 2,471 affected children and 3,018 unaffected parents. In order to select the children and parents of European ancestry, we performed multi-dimensional scaling (MDS) in PLINK [29] with the samples from the 1000 Genomes Project as reference, and matched based on the first 10 dimensions. To reduce genotyping errors, variants with HWE p-values of $\leq 1\times10^{-3}$ in the controls were removed. We performed Fisher's Exact Test on the allele counts for common (>5%), low-frequency (1-5%) and rare (<1%) variants to ensure that there is no global inflation and that we had a homogeneous European population for our study. The variants were then annotated using Variant Effect Predictor [30].

Exome sequencing quality control and analyses

Whole-exome sequencing on the Middle Eastern individuals with autism and intellectual disability were performed at the Broad Institute [4]. Given the expected higher rate of consanguinity in these families, we selected 1 unaffected individual (parent or sibling) from each family to evaluate the allele frequencies of the variants. Variants that passed the standard GATK filters, had read depth of ≥10, Phred quality score of ≥30, and alternate heterozygous reads of 30-70% or alternate homozygous reads of >90% were kept for analyses. In addition, variants with >10% missing genotypes after filtering were discarded as these are likely to be found in regions of the exome with poor coverage and can result in inaccurate estimates of the allele frequencies.

RAFT test statistic

Let a denote the minor or non-reference allele while A is the reference or major allele. Let the probability of the expected number of individuals with the homozygous minor allele be $P(aa)$. There are a few ways to estimate $P(aa)$, such as using the heterozygotes in cases and controls for the estimation. One approach for estimating $P(aa)$ is to calculate $P(a|case)$ and $P(a|control)$ separately and we can estimate:

$$P(aa|case) = P(a|case) \times P(a|case)$$

$$P(aa|control) = P(a|control) \times P(a|control)$$

An alternative approach is to estimate $P(aa) = P(a) \times P(a)$ by using the observed allele counts for the heterozygotes in cases and controls, which is the maximum likelihood estimate for the allele frequency:

$$P(a) = \frac{n_{case_{Aa}} + n_{control_{Aa}}}{n_{case_{Aa}} + 2n_{case_{AA}} + n_{control_{Aa}} + 2n_{control_{AA}}}$$

Then we performed an expectation maximization (EM) step to calculate the probability of observing n_{aa} number of alternate homozygotes in the controls, $P(n_{aa}|control, \gamma = \gamma_{controls})$ and the probability of observing $n_{\overline{aa}}$ number of heterozygotes and reference homozygotes in the controls, $P(n_{\overline{aa}} \,|\, control, \gamma = \gamma_{controls})$ given a genotypic risk ratio in the controls ($\gamma_{controls}$). The binomial probability for such an observation can be calculated using $P(n_{aa}|control, \gamma_{controls}) \times P(n_{\overline{aa}} \,|\, control, \gamma_{controls})$, which is equivalent to $\frac{[\gamma_{controls} \times P(aa)]^{n_{aa}}[1-\gamma_{controls} \times P(aa)]^{n_{\overline{aa}}}}{\gamma_{controls} \times P(aa) + (1-P(aa))}$. Essentially, if there are variants with excessive homozygosity in the controls and deviate from the squared of the allele frequencies (which are

the expected probabilities for the homozygotes), this can be an indication of a common copy number polymorphism unmasking a rare variant in the controls, or variants in poorly covered regions with inaccurate variant calls, or misalignments with the human reference sequence. In such instances, $\gamma_{controls}$ will be greater than 1, otherwise $\gamma_{controls}$ will be equal to 1.

Next, we performed a second EM step to estimate the equivalent γ_{cases} in the cases and the formulation for the RAFT test statistic is a log-likelihood ratio between the observed and expected probabilities of recessives in the cases with respect to the controls: $RAFT = log_{10}\left[\frac{P(n_{aa}|case,\gamma=\gamma_{\text{cases}})P(n_{\overline{aa}}|case,\gamma=\gamma_{cases})}{P(n_{aa}|control,\gamma=\gamma_{controls})P(n_{\overline{aa}}|control,\gamma=\gamma_{controls})}\right]$. This evaluates any homozygosity in the cases in excess of the homozygosity found in the controls. In addition, we multiplied the probability by a global correction factor *f* to account for any excess of homozygosity in the exome (see Supplementary Methods). For instance, in consanguineous populations, we expect more rare homozygous variants than in out-bred populations, so the correction factor will result in a global reduction in the significance of a single rare homozygous variant in a consanguineous family.

Mutagenesis and molecular cloning in yeast

The human *DHCR24* cDNA clones (BC004375 and BC01169) were ordered from Thermo Scientific and Sanger sequencing was performed on the cDNA clones to verify the sequences. We performed *EcoRI* and *XhoI* restriction enzyme digests on the 2 cDNA clones, as well as *EcoRI* digest on the FB1533 or p426 GAL vector to ligate the *DHCR24* cDNA clones into the FB1533 vector. This resulted in a final plasmid that is 8.8kb long: 6.4kb for the vector and 2.4kb for the cDNA. We performed transformation using TOP10 competent cells from Life Technologies and cultured the *E. coli* for mutagenesis, which was done using the Agilent

QuikChange II site-directed mutagenesis kit. The primers used for mutagenesis are listed in Table 3.2 and Sanger sequencing was performed to confirm the mutants. The mutant plasmids were then transformed into the FY1856 yeast strain lacking the URA3 gene and colonies were selected on media lacking uracil.

Yeast cell lysis and cholesterol synthesis biochemical assay

Yeast cells were grown overnight to saturation (OD ~1.5), spun down and washed with 5mM Tris/HCl (pH 7.5) and 50 mM NaCl buffer, before re-suspending in 500µl of NaCl buffer. The cells were flash freezed and glass beads (0.40 to 0.60mm) from Jencons Scientific were added. Cell lysis was performed using a Beadbeater for 1 minute at 4°C and the resultant lysates were placed on ice for 2 minutes before extracting the supernatant. Using a previously described protocol, we incubated 25µl of the supernatant with 225µl of assay buffer comprising of desmosterol for 4 hours at 37°C before performing the sterol analyses [20].

Sterol analyses

The assay mixes were sent overnight on dry ice to the Clinical Mass Spectrometry Laboratory at Kennedy Krieger Institute to measure the amounts of desmosterol, cholesterol and ergosterol.

ACKNOWLEDGEMENTS

The authors would like to thank the NHLBI GO Exome Sequencing Project and its ongoing studies which produced and provided exome variant calls for comparison: the Lung GO Sequencing Project (HL-102923), the WHI Sequencing Project (HL-102924), the Broad GO Sequencing Project (HL-102925), the Seattle GO Sequencing Project (HL-102926) and the Heart GO Sequencing Project (HL-103010).

Table 3.2: Primer sequences used for mutagenesis in *DHCR24*

Mutant	Primer Names	Primer Sequences (5' to 3')	Variant
1	g1522a_sense	cttccccgaggtgtacaacaagatctgcaaggc	p. D508N or c.1522G>A
	g1522a_antisense	gccttgcagatcttgttgtacacctcggggaag	
2	g1353c_sense	acactttgaagccagctcctgcatgaggcag	p. R451S or c.1353G>C
	g1353c_antisense	ctgcctcatgcaggagctggcttcaaagtgt	
3	a1301g_sense	aaatgaggcagagctctgcatcgacattggagcat	p.Y434C or c.1301A>G
	a1301g_antisense	atgctccaatgtcgatgcagagctctgcctcattt	
4	g1140c_sense	agctgtacgagcaccaccacgtggtgc	p.Q380H or c.1140G>C
	g1140c_antisense	gcaccacgtggtggtgctcgtacagct	
5	g1135a_sense	tgcgcaagctgtacaagcagcaccacgtg	p.E379K or c.1135G>A
	g1135a_antisense	cacgtggtgctgcttgtacagcttgcgca	
6	a815c_sense	ggaagggctgctctcctccctggatgagg	p.Y272S or c.815A>C
	a815c_antisense	cctcatccagggaggagagcagcccttcc	
7	g737a_sense	cgagccagtgcagggcctggagg	p.R246Q or c.737G>A
	g737a_antisense	cctccaggccctgcactggctcg	
8	c736t_sense	ttcgagccagtgtggggcctggagg	p.R246W or c.736C>T
	c736t_antisense	cctccaggccccacactggctcgaa	
9	c711a_sense	gcatcatccctgccaagaagtaagtcaagctgcgt	p.Y237X or c.711C>A
	c711a_antisense	acgcagcttgacttacttcttggcagggatgatgc	

Table 3.2: Primer sequences used for mutagenesis in *DHCR24* (Continued)

Mutant	Primer Names	Primer Sequences (5' to 3')	Variant
10	g689a_sense	gccgctgagatccacatcatccctgcc	p.R230H or c.689G>A
	g689a_antisense	ggcagggatgatgtggatctcagcggc	
11	g616a_sense	gcgatgcactccgtccaaaaactcagacctgtt	p.E206K or c.616G>A
	g616a_antisense	aacaggtctgagtttttggacggagtgcatcgc	
12	c566g_sense	caacacatctgcactggttacgagctggtcctg	p.A189G or c.566C>G
	c566g_antisense	caggaccagctcgtaaccagtgcagatgtgttg	
13	c519g_sense	catgggcacaggcatggagtcatcatccca	p.I173M or c.519C>G
	c519g_antisense	tgggatgatgactccatgcctgtgcccatg	
14	c403t_sense	attgtccgtgtggagtccttggtgaccatgg	p.P135S or c.403C>T
	c403t_antisense	ccatggtcaccaaggactccacacggacaat	
15	c394t_sense	ccaagaaacagattgtctgtgtggagcccttggtg	p.R132C or c.394C>T
	c394t_antisense	caccaagggctccacacagacaatctgtttcttgg	
16	g236a_sense	ccagaagcaggtgcaggaatggaaggagc	p.R79Q or c.236G>A
	g236a_antisense	gctccttccattcctgcacctgcttctgg	
17	a881c_sense	agagcccagcaagctgactagcattggcaattact	p.N294T or c.881A>C
	a881c_antisense	agtaattgccaatgctagtcagcttgctgggctct	
18	g918t_sense	caattactacaagccgtggttctttaatcatgtggagaactat	p.K306N or c.918G>T
	g918t_antisense	atagttctccacatgattaaagaaccacggcttgtagtaattg	
19	c731t_sense	ctgcgtttcgagctagtgcggggcctg	p.P244L or c.731C>T

Table 3.2: Primer sequences used for mutagenesis in *DHCR24* (Continued)

Mutant	Primer Names	Primer Sequences (5' to 3')	Variant
	c731t_antisense	caggccccgcactagctcgaaacgcag	
20	g308a_sense	gctcactgtctcactacatgtcgggaagtacaaga	p.R103H or c.G308A
	g308a_antisense	tcttgtacttcccgacatgtagtgagacagtgagc	
21	a1412c_sense	catggcttccagatgctgtctgccgactgc	p.Y471S or c.A1412C
	a1412c_antisense	gcagtcggcagacagcatctggaagccatg	
22	g1433a_sense	gactgctacatgaaccaggaggagttctgggag	p.R478Q or c.1433G>A
	g1433a_antisense	ctcccagaactcctcctggttcatgtagcagtc	
23	c108t_sense	tgggtgttcgtgtgtctcttcctcctgcc	p.C36X or c.108C>T
	c108t_antisense	ggcaggaggaagagacacacgaacaccca	
24	c1328t_sense	ggagcatatggggagctgcgtgtgaaacacttt	p.P443L or c.1328C>T
	c1328t_antisense	aaagtgtttcacacgcagctccccatatgctcc	

SUPPLEMENTARY MATERIALS AND METHODS

Supplementary Methods

Evaluating the RAFT test statistic on a simulated Finnish dataset

To evaluate the distribution of our test statistic, we obtained a whole-exome sequencing dataset comprising of 3,000 individuals of Finnish ancestry and 3,000 non-Finnish Europeans (NFEs) (Lim et al., unpublished). There are a total of 590,003 coding variants in 3,000 Finns and 3,000 NFEs from whole exome sequencing data, and for each variant, we randomly generated 5,000 cases and 5,000 controls using the allele frequencies derived from only the Finns. When we ran RAFT on the simulated Finnish dataset with 5,000 cases and 5,000 controls, we found that across all allele frequency bins (common ≥5%, low-frequency 1-5% or rare ≤1%), we obtained similar numbers of observed compared to expected variants in the various p-value bins and the ratios of the expected compared to observed are approximately 1 or less than 1 (Table 3.3).

Evaluating the RAFT test statistic on a simulated Finnish and NFE dataset with substructure

There are 2 confounding factors in applying RAFT – first, population stratification where the cases are from an ethnically different population compared to the controls, can result in deviation from HWE. However, similar to existing genome-wide association studies, this can be easily detected using existing methods such as genomic control to identify if the cases are well-matched to the controls in terms of ancestry. The second confounding factor is population substructure where the cases and controls are equally sampled from two or more heterogeneous populations. For instance, if the cases and controls are derived from Finns and NFEs, this can result in the Wahlund effect where there is excessive homozygosity and reduced heterozygosity. Unlike

direct case-control comparisons, such heterogeneity can inflate the statistic above. To further evaluate the effect of population sub-structure on RAFT (where the cases and controls are well-matched for ancestry but contain ethnically different subpopulations in both cases and controls), we randomly generated 5,000 cases and 5,000 controls using equal proportions of Finns and NFEs in the cases and controls. However, when we ran the test statistic on the simulated data, we observed an unusually high inflation among the common variants (Table 3.4). Even in a study in which cases and controls are perfectly matched exome-wide, departures from HWE may occur at sites if the population is not of homogeneous ancestry (such as the Wahlund effect). As such, careful analyses have to be performed to ensure that there is no excessive substructure in the cases and controls when applying RAFT.

Table 3.3: Distribution of variants from running RAFT on the simulated Finnish dataset

The observed median number of variants and the lower and upper values from 100 simulations are shown in the brackets.

Number of common (≥5%) variants = 26,368					
	$P \leq 1 \times 10^{-5}$	$P \leq 1 \times 10^{-4}$	$P \leq 1 \times 10^{-3}$	$P \leq 1 \times 10^{-2}$	$P \leq 1 \times 10^{-1}$
Observed	0 [0-3]	3 [0-9]	17 [7-29]	117 [96-139]	765 [704-844]
Expected	0.26	2.6	26	264	2637
Ratio (Obs/Exp)	0 [0-11.4]	1.14 [0-3.41]	0.64 [0.27-1.1]	0.44 [0.36-0.53]	0.29 [0.27-0.32]
Number of low-freq (1-5%) variants = 15,665					
	$P \leq 1 \times 10^{-5}$	$P \leq 1 \times 10^{-4}$	$P \leq 1 \times 10^{-3}$	$P \leq 1 \times 10^{-2}$	$P \leq 1 \times 10^{-1}$
Observed	0 [0-1]	0 [0-3]	5 [0-12]	91	682
Expected	0.16	1.57	16	157	1567
Ratio (Obs/Exp)	0 [0-6.38]	0 [0-1.92]	0.32 [0-0.76]	0.29 [0.17-0.39]	0.28 [0.25-0.31]
Number of rare (≤1%) variants = 161,291					
	$P \leq 1 \times 10^{-5}$	$P \leq 1 \times 10^{-4}$	$P \leq 1 \times 10^{-3}$	$P \leq 1 \times 10^{-2}$	$P \leq 1 \times 10^{-1}$
Observed	0 [0-2]	6 [1-17]	43 [28-59]	196 [162-232]	881 [785-948]
Expected	1.61	16	161	1613	16,129
Ratio (Obs/Exp)	0 [0-1.24]	0.37 [0.06-1.05]	0.27 [0.17-0.37]	0.12 [0.1-0.14]	0.05 [0.05-0.06]

Table 3.4: Distribution of variants from running RAFT on the simulated Finnish and NFE dataset

The observed median number of variants and the lower and upper values from 100 simulations are shown in the brackets.

Number of common (≥5%) variants = 25,184					
	$P \leq 1 \times 10^{-5}$	$P \leq 1 \times 10^{-4}$	$P \leq 1 \times 10^{-3}$	$P \leq 1 \times 10^{-2}$	$P \leq 1 \times 10^{-1}$
Observed	3 [0-7]	10 [4-17]	44 [33-65]	217 [189-242]	1101 [1041-1163]
Expected	0.25	2.52	25	252	2518
Ratio (Obs/Exp)	11.9 [0-27.8]	3.97 [1.59-6.75]	1.75 [1.31-2.58]	0.86 [0.75-0.96]	0.44 [0.41-0.46]
Number of low-freq (1-5%) variants = 15,165					
	$P \leq 1 \times 10^{-5}$	$P \leq 1 \times 10^{-4}$	$P \leq 1 \times 10^{-3}$	$P \leq 1 \times 10^{-2}$	$P \leq 1 \times 10^{-1}$
Observed	0.5 [0-4]	4 [0-9]	18 [8-29]	108 [87-130]	684 [609-748]
Expected	0.15	1.52	15	152	1517
Ratio (Obs/Exp)	3.3 [0-26.4]	2.64 [0-5.93]	1.19 [0.53-1.91]	0.71 [0.57-0.86]	0.45 [0.4-0.49]
Number of rare (≤1%) variants = 270,458					
	$P \leq 1 \times 10^{-5}$	$P \leq 1 \times 10^{-4}$	$P \leq 1 \times 10^{-3}$	$P \leq 1 \times 10^{-2}$	$P \leq 1 \times 10^{-1}$
Observed	1 [0-3]	28 [18-47]	120.5 [94-143]	419 [379-485]	1561.5 [1462-1650]
Expected	2.7	27	270	2705	27046
Ratio (Obs/Exp)	0.37 [0-1.11]	1.04 [0.67-1.74]	0.45 [0.35-0.53]	0.15 [0.14-0.18]	0.06 [0.05-0.06]

Calculating correction factor to detect substructure

To estimate the amount of substructure or homozygosity by descent, we fitted a regression model on all coding variants with the intercept set to 0, where q is the allele frequency of the alternate allele:

$$\frac{Number\ of\ homozygotes}{Number\ of\ individuals} = \beta_1\, q + \beta_2\, q^2$$

Using the whole-exome sequencing data for the 3,000 NFEs, we estimated the parameters:

$$\beta_1 = 0.00898$$

$$\beta_2 = 0.991$$

Using the whole-exome sequencing data for the 3,000 Finns, we estimated the parameters:

$$\beta_1 = 0.00675$$

$$\beta_2 = 0.993$$

Using the whole-exome sequencing data for the 207 Middle Eastern individuals, we estimated the parameters:

$$\beta_1 = 0.0457$$

$$\beta_2 = 0.954$$

Using the whole-exome sequencing data for the combined Finn and NFE individuals to simulate a population with substructure, we estimated the parameters:

$$\beta_1 = 0.0121$$

$$\beta_2 = 0.988$$

As a comparison, we compared the expected probabilities for homozygotes versus the calculated probabilities using the fitted models across the 3 populations (Table 3.5).

Table 3.5: Comparison of homozygosity by descent or population substructure

The expected probabilities, as well as calculated probabilities after correcting for homozygosity by descent or population substructure across the non-Finnish European (NFE), Finnish and Middle Eastern populations across different allele frequency ranges.

Allele Freq	**Expected**	**NFE Calculated**	**NFE Inflation (Calculated/ Expected)**	**Finn Calculated**	**Finn Inflation (Calculated/ Expected)**	**Middle Eastern Calculated**	**Middle Eastern Inflation (Calculated/ Expected)**
0.1%	1×10^{-6}	9.97×10^{-6}	9.97	7.74×10^{-6}	7.74	4.67×10^{-5}	46.65
0.5%	2.5×10^{-5}	6.97×10^{-5}	2.79	5.86×10^{-5}	2.34	2.52×10^{-4}	10.09
1%	1×10^{-4}	1.89×10^{-4}	1.89	1.67×10^{-4}	1.67	5.52×10^{-4}	5.52
5%	2.5×10^{-3}	2.92×10^{-3}	1.17	2.82×10^{-3}	1.13	4.67×10^{-3}	1.87
10%	0.01	0.011	1.08	0.011	1.06	0.014	1.41
20%	0.04	0.041	1.04	0.041	1.03	0.047	1.18
30%	0.09	0.092	1.02	0.091	1.02	0.1	1.11
40%	0.16	0.16	1.01	0.16	1.01	0.17	1.07
50%	0.25	0.25	1	0.25	1	0.26	1.05

Calculating significance for the 3 homozygous non-synonymous observations in *BCKDK*

We assumed a non-synonymous composite allele frequency of 0.0024 for *BCKDK* in 207 control Middle Eastern individuals (1 non-synonymous heterozygote in 414 chromosomes), and that there was 1 heterozygote and 3 homozygotes observed in 208 Middle Eastern cases with autism. The probability of observing a homozygote adjusted for homozygosity by descent or population substructure is 0.00012, representing an inflation of 19.3 from the expected probability. The RAFT LOD score for these 3 homozygous observations is 4.99, $P = 1.64\times10^{-6}$. Given that there are 3 additional affected siblings who are homozygous and 9 unaffected siblings who are not homozygous across these families, the final p-value for this observation is $1.64\times10^{-6} \times (0.25)^3 \times (0.75)^9 = 1.92\times10^{-9}$.

Calculating significance for 2 homozygous non-synonymous observations in *DHCR24* in the Middle Eastern population

For the 2 Middle Eastern homozygotes, we assumed a non-synonymous composite allele frequency of 0.0048 (2 heterozygotes in 414 chromosomes), and that there were 2 heterozygotes and 2 homozygotes observed in 207 Middle Eastern cases with autism or intellectual disability. The probability of observing a homozygote adjusted for homozygosity by descent or population substructure is 0.00025, representing an inflation of 10.7 from the expected probability. The RAFT LOD score for these 2 homozygous observations is 2.35, $P = 1\times10^{-3}$. Given that there are 3 additional affected siblings who are homozygous and 9 unaffected siblings who are not homozygous, the final p-value for this observation is $1\times10^{-3} \times (0.25)^3 \times (0.75)^9 = 1.18\times10^{-6}$.

Calculating significance for 1 homozygous non-synonymous observations in *DHCR24* in the European population

For the 1 European homozygote, we assumed a non-synonymous composite allele frequency of 0.00442 (38 heterozygotes in 8,600 chromosomes), and that there were 5 heterozygotes and 1 homozygote observed in 1,063 European cases with autism. The probability of observing a homozygote adjusted for homozygosity by descent or population substructure is 5.22×10^{-5}, representing an inflation of 2.7 from the expected probability. The RAFT LOD score for this homozygous observation is 0.84, P = 0.049. Given that there are 2 additional affected siblings who are homozygous, the final p-value for this observation is $0.049 \times (0.25)^2 = 3.08\times10^{-3}$.

Calculating significance for combined observations in *DHCR24* using meta-analysis

In order to obtain a chi-squared statistic with 1 degree of freedom, we performed meta-analysis across the Middle Eastern and European populations to evaluate the significance of the 3 homozygotes in *DHCR24* in both populations. We performed the expectation maximization step to obtain the genotype relative risk that maximizes the sum of the log likelihood function for both populations. That is:

$$RAFT = log_{10}\left[\frac{\begin{matrix}P\left(n_{aa\_ME}\middle|case,\gamma=\gamma_{\text{cases}}\right)P\left(n_{\overline{aa\_ME}}\middle|case,\gamma=\gamma_{cases}\right)\\ P\left(n_{aa\_EA}\middle|case,\gamma=\gamma_{\text{cases}}\right)P\left(n_{\overline{aa\_EA}}\middle|case,\gamma=\gamma_{cases}\right)\end{matrix}}{\begin{matrix}P\left(n_{aa\_ME}\middle|control,\gamma=\gamma_{controls}\right)P\left(n_{\overline{aa\_ME}}\middle|control,\gamma=\gamma_{controls}\right)\\ P\left(n_{aa\_EA}\middle|control,\gamma=\gamma_{controls}\right)P\left(n_{\overline{aa\_EA}}\middle|control,\gamma=\gamma_{controls}\right)\end{matrix}}\right]$$

where

$P\left(n_{aa\_ME}\middle|case,\gamma=\gamma_{\text{cases}}\right)$ is the probability of homozygotes in the Middle Eastern (ME) cases,

$P\left(n_{\overline{aa\_ME}}\middle|case,\gamma=\gamma_{cases}\right)$ is the probability of non-homozygotes in the Middle Eastern cases,

$P\left(n_{aa\_EA}\middle|case,\gamma=\gamma_{\text{cases}}\right)$ is the probability of homozygotes in the European (EA) cases,

$P\left(n_{\overline{aa\_EA}} \middle| case, \gamma = \gamma_{cases}\right)$ is the probability of non-homozygotes in the European cases,

$P\left(n_{aa\_ME} \middle| control, \gamma = \gamma_{controls}\right)$ is the probability of homozygotes in the Middle Eastern controls,

$P\left(n_{\overline{aa\_ME}} \middle| control, \gamma = \gamma_{controls}\right)$ is the probability of non-homozygotes in the Middle Eastern controls,

$P\left(n_{aa\_EA} \middle| control, \gamma = \gamma_{controls}\right)$ is the probability of homozygotes in the European controls,

$P\left(n_{\overline{aa\_EA}} \middle| control, \gamma = \gamma_{controls}\right)$ is the probability of the non-homozygotes in the European controls.

The allele frequencies and inflation correction factors used are the same as mentioned above for the Middle Eastern and European populations. The meta-analyzed RAFT LOD score for these 3 homozygous observations is 4.22, $P = 1.04\times10^{-5}$. Given that there are 5 additional affected siblings who are homozygous and 9 unaffected siblings who are not homozygous, the final p-value for this observation is $1.04\times10^{-5} \times (0.25)^5 \times (0.75)^9 = 7.64\times10^{-10}$.

REFERENCES

1. Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, et al. (2010) Exome sequencing identifies the cause of a mendelian disorder. Nat Genet 42: 30-35.

2. Chahrour MH, Yu TW, Lim ET, Ataman B, Coulter ME, et al. (2012) Whole-exome sequencing and homozygosity analysis implicate depolarization-regulated neuronal genes in autism. PLoS Genet 8: e1002635.

3. Lim ET, Raychaudhuri S, Sanders SJ, Stevens C, Sabo A, et al. (2013) Rare complete knockouts in humans: population distribution and significant role in autism spectrum disorders. Neuron 77: 235-242.

4. Yu TW, Chahrour MH, Coulter ME, Jiralerspong S, Okamura-Ikeda K, et al. (2013) Using whole-exome sequencing to identify inherited causes of autism. Neuron 77: 259-273.

5. Keller MC, Simonson MA, Ripke S, Neale BM, Gejman PV, et al. (2012) Runs of homozygosity implicate autozygosity as a schizophrenia risk factor. PLoS Genet 8: e1002656.

6. Morrow EM, Yoo SY, Flavell SW, Kim TK, Lin Y, et al. (2008) Identifying autism loci and genes by tracing recent shared ancestry. Science 321: 218-223.

7. Styrkarsdottir U, Thorleifsson G, Sulem P, Gudbjartsson DF, Sigurdsson A, et al. (2013) Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits. Nature 497: 517-520.

8. Jonsson T, Atwal JK, Steinberg S, Snaedal J, Jonsson PV, et al. (2012) A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. Nature 488: 96-99.

9. Gamsiz ED, Viscidi EW, Frederick AM, Nagpal S, Sanders SJ, et al. (2013) Intellectual disability is associated with increased runs of homozygosity in simplex autism. Am J Hum Genet 93: 103-109.

10. Novarino G, El-Fishawy P, Kayserili H, Meguid NA, Scott EM, et al. (2012) Mutations in BCKD-kinase lead to a potentially treatable form of autism with epilepsy. Science 338: 394-397.

11. Wittke-Thompson JK, Pluzhnikov A, Cox NJ (2005) Rational inferences about departures from Hardy-Weinberg equilibrium. Am J Hum Genet 76: 967-986.

12. Song K, Elston RC (2006) A powerful method of combining measures of association and Hardy-Weinberg disequilibrium for fine-mapping in case-control studies. Stat Med 25: 105-126.

13. Curtis D (2013) Consideration of plausible genetic architectures for schizophrenia and implications for analytic approaches in the era of next generation sequencing. Psychiatr Genet 23: 1-10.

14. Ronemus M, Iossifov I, Levy D, Wigler M (2014) The role of de novo mutations in the genetics of autism spectrum disorders. Nat Rev Genet 15: 133-141.

15. Zolotushko J, Flusser H, Markus B, Shelef I, Langer Y, et al. (2011) The desmosterolosis phenotype: spasticity, microcephaly and micrognathia with agenesis of corpus callosum and loss of white matter. Eur J Hum Genet 19: 942-946.

16. Schaaf CP, Koster J, Katsonis P, Kratz L, Shchelochkov OA, et al. (2011) Desmosterolosis-phenotypic and molecular characterization of a third case and review of the literature. Am J Med Genet A 155A: 1597-1604.

17. Andersson HC, Kratz L, Kelley R (2002) Desmosterolosis presenting with multiple congenital anomalies and profound developmental delay. Am J Med Genet 113: 315-319.

18. FitzPatrick DR, Keeling JW, Evans MJ, Kan AE, Bell JE, et al. (1998) Clinical phenotype of desmosterolosis. Am J Med Genet 75: 145-152.

19. Zerenturk EJ, Sharpe LJ, Ikonen E, Brown AJ (2013) Desmosterol and DHCR24: unexpected new directions for a terminal step in cholesterol synthesis. Prog Lipid Res 52: 666-680.

20. Waterham HR, Koster J, Romeijn GJ, Hennekam RC, Vreken P, et al. (2001) Mutations in the 3beta-hydroxysterol Delta24-reductase gene cause desmosterolosis, an autosomal recessive disorder of cholesterol biosynthesis. Am J Hum Genet 69: 685-694.

21. Jira P (2013) Cholesterol metabolism deficiency. Handb Clin Neurol 113: 1845-1850.

22. Kaminsky EB, Kaul V, Paschall J, Church DM, Bunke B, et al. (2011) An evidence-based approach to establish the functional and clinical significance of copy number variants in intellectual and developmental disabilities. Genet Med 13: 777-784.

23. Greeve I, Hermans-Borgmeyer I, Brellinger C, Kasper D, Gomez-Isla T, et al. (2000) The human DIMINUTO/DWARF1 homolog seladin-1 confers resistance to Alzheimer's disease-associated neurodegeneration and oxidative stress. J Neurosci 20: 7345-7352.

24. Iivonen S, Hiltunen M, Alafuzoff I, Mannermaa A, Kerokoski P, et al. (2002) Seladin-1 transcription is linked to neuronal degeneration in Alzheimer's disease. Neuroscience 113: 301-310.

25. Wisniewski T, Newman K, Javitt NB (2013) Alzheimer's disease: brain desmosterol levels. J Alzheimers Dis 33: 881-888.

26. Sharpe LJ, Wong J, Garner B, Halliday GM, Brown AJ (2012) Is seladin-1 really a selective Alzheimer's disease indicator? J Alzheimers Dis 30: 35-39.

27. Tierney E, Bukelis I, Thompson RE, Ahmed K, Aneja A, et al. (2006) Abnormalities of cholesterol metabolism in autism spectrum disorders. Am J Med Genet B Neuropsychiatr Genet 141B: 666-668.

28. Buchovecky CM, Turley SD, Brown HM, Kyle SM, McDonald JG, et al. (2013) A suppressor screen in Mecp2 mutant mice implicates cholesterol metabolism in Rett syndrome. Nat Genet 45: 1013-1020.

29. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81: 559-575.

30. McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, et al. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics 26: 2069-2070.

CHAPTER 4

Discovery of medically important loss-of-function variants by utilizing the Finnish founding bottleneck

Discovery of medically important loss-of-function variants by utilizing the Finnish founding bottleneck

Elaine T. Lim[1-4], Peter Würtz[5,6], Aki S. Havulinna[6], Priit Palta[5,7], Taru Tukiainen[1-3], Karola Rehnström[7], Tõnu Esko[2,3,8,9], Reedik Mägi[8], Michael Inouye[10], Tuuli Lappalainen[11,12], Xueling Sim[13], Alisa Manning[2], Claes Ladenvall[5,14], Suzannah Bumpstead[7], Eija Hämäläinen[5,7], Kristiina Aalto[15], Mikael Maksimow[15], Marko Salmi[16], Stefan Blankenberg[17,18], Diego Ardissino[19], Svati Shah[20], Benjamin Horne[21], Ruth McPherson[22], Gerald K. Hovingh[23], Muredach P. Reilly[24], Hugh Watkins[25], Anuj Goel[25], Martin Farrall[25], Domenico Girelli[26], Alex P. Reiner[27], Nathan O. Stitziel[28], Sekar Kathiresan[29], Stacey Gabriel[2], Jeffrey C. Barrett[7], Terho Lehtimäki[30], Markku Laakso[31], Leif Groop[5,14], Jaakko Kaprio[5,32,33], Markus Perola[5], Mark I. McCarthy[34-36], Michael Boehnke[13], David M. Altshuler[2,3], Cecilia M. Lindgren[1,2,37], Joel N. Hirschhorn[3,9,38], Andres Metspalu[8], Nelson B. Freimer[39], Tanja Zeller[17,18], Sirpa Jalkanen[16], Seppo Koskinen[40], Olli Raitakari[41,42], Richard Durbin[7], Daniel G. MacArthur[1-3], Veikko Salomaa[6], Samuli Ripatti[5-7,32,43], Mark J. Daly[1-3*¶], Aarno Palotie[1,2,5,44*¶] for the Sequencing Initiative Suomi (SISu) Project

1 Analytic and Translational Genetics Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA.

2 Program in Medical and Population Genetics, Broad Institute, Cambridge, MA, USA.

3 Department of Genetics, Harvard Medical School, Boston, MA, USA.

4 Program in Biological and Biomedical Sciences, Harvard Medical School, Boston, MA, USA.

5 Institute for Molecular Medicine Finland, University of Helsinki, Helsinki, Finland.

6 Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland.

7 Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK.

8 Estonian Genome Center, University of Tartu, Tartu, Estonia.

9 Divisions of Endocrinology and Genetics and Center for Basic and Translational Obesity Research, Children's Hospital Boston, Boston, MA.

10 Medical Systems Biology, Department of Pathology and Department of Microbiology & Immunology, The University of Melbourne, Parkville, Victoria 3010, Australia.

11 Department of Genetics, Stanford University. 365 Lasuen Street, Littlefield Center, Stanford, CA.

12 Stanford Center for Computational, Evolutionary and Human Genomics. 365 Lasuen Street, Littlefield Center, Stanford, CA.

13 Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, USA.

14 Lund University Diabetes Center, Department of Clinical Sciences, Diabetes & Endocrinology, Skåne University Hospital, Lund University, Malmö, Sweden.

15 MediCity, University of Turku, Turku, Finland.

16 Department of Medical Microbiology and Immunology, University of Turku and National Institute for Health and Welfare, Turku, Finland.

17 University Heart Centre Hamburg, Clinic for General and Interventional Cardiology, Hamburg, Germany.

18 DZHK (German Centre for Cardiovascular Research), partner site Hamburg/Kiel/Lübeck, Hamburg, Germany.

19 Division of Cardiology, Azienda Ospedaliero-Universitaria di Parma, Parma, Italy.

20 Department of Medicine, Duke University Medical Center, Durham, North Carolina.

21 Intermountain Heart Institute, Intermountain Medical Center, Salt Lake City, Utah, USA.

22 Division of Cardiology, University of Ottawa Heart Institute, Ottawa, ON, Canada.

23 Department of Vascular Medicine, Academic Medical Center, Amsterdam, The Netherlands.

24 Cardiovascular Institute, Perelman School of Medicine at the University of Pennsylvania, Philadelphia, PA, USA.

25 Division of Cardiovascular Medicine, Radcliffe Department of Medicine, The Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK.

26 University of Verona School of Medicine, Department of Medicine, Verona, Italy.

27 Department of Epidemiology, University of Washington, Seattle, Washington, USA.

28 Cardiovascular Division, Department of Medicine, Washington University School of Medicine, St. Louis, MO, USA.

29 Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts, USA.

30 Department of Clinical Chemistry, Fimlab Laboratories, University of Tampere School of Medicine, Tampere, Finland.

31 Department of Medicine, University of Eastern Finland, Kuopio, Finland.

32 University of Helsinki, Hjelt Institute , Dept of Public Health, P.O.Box 41 Mannerheimintie 172, 00014 Helsinki, Finland.

33 National Institute for Health and Welfare, Dept of Mental Health and Substance Abuse Services, P.O. Box 30 (Mannerheimintie 166), 00300 Helsinki, Finland.

34 Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Old Road, Headington, Oxford, OX3 7LJ UK.

35 Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK.

36 Oxford NIHR Biomedical Research Centre, Churchill Hospital, Old Road, Headington, Oxford, OX3 7LJ UK.

37 Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom.

38 Division of Endocrinology, Department of Medicine, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA.

39 University of California Los Angeles Center for Neurobehavioral Genetics, Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Los Angeles, CA, USA.

40 Department of Health, Functional Capacity and Welfare, National Institute for Health and Welfare, Helsinki, Finland.

41 Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Finland.

42 Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Turku, Finland.

43 Department of Biometry, Hjelt Institute, University of Helsinki.

44 Psychiatric & Neurodevelopmental Genetics Unit, Department of Psychiatry, Massachusetts General Hospital, Boston, MA, USA.

[*] These authors contributed equally.

[¶] Corresponding authors: MJD – mjdaly@atgu.mgh.harvard.edu, AP - aarno@broadinstitute.org

Manuscript in review.

ABSTRACT

Exome sequencing studies in complex diseases are challenged by the allelic heterogeneity, large number and modest effect sizes of associated variants on disease risk and the presence of large numbers of neutral variants, even in phenotypically relevant genes. Isolated populations with recent bottlenecks offer advantages for studying rare variants in complex diseases as they have deleterious variants that are present at higher frequencies as well as a substantial reduction in rare neutral variation. To explore the potential of the Finnish founder population for studying low-frequency (0.5-5%) variants in complex diseases, we compared exome sequence data on 3,000 Finns to the same number of non-Finnish Europeans (NFEs) and used several well-characterized population cohorts in a reverse genetics approach to genotype 83 low-frequency loss-of-function variants in 36,262 Finns. Using a deep set of quantitative traits collected on these cohorts, we identified splice variants in *LPA* that lowered plasma lipoprotein(a) levels as well as novel associations with circulating D-dimer levels, galectin-3 levels, triglycerides and systolic blood pressure. Through accessing the national medical records of these participants, we could evaluate the *LPA* finding via Mendelian randomization and confirm that these splice variants confer protection from cardiovascular disease (OR = 0.84, $P = 3\times10^{-4}$), demonstrating for the first time that inhibition of *LPA* may have therapeutic efficacy. More generally, this study articulates substantial advantages for studying the role of rare variation in complex phenotypes in Finland - combining a unique population genetic history with data from large population cohorts and centralized research access to medical records and National Health Registers.

INTRODUCTION

After widespread success with genome-wide association studies (GWAS) of common variants, several studies have recently begun to identify rare (with <0.5% allele frequency) and low-frequency (0.5-5%) variants in complex diseases and traits such as triglycerides [1], insulin processing [2], bone mineral density [3], Alzheimer's disease [4], impulsivity [5], and prostate cancer [6], some of which confer protection from disease [4]. Protective loss of function variants that can be tolerated in a homozygote state in humans are of particular interest as potential safe targets for therapeutic inhibition. Interestingly, many of these studies that have discovered rare and low-frequency variants have benefited from the use of isolated populations that have undergone bottlenecks resulting in frequency enrichment of the associated variants. In contrast to the large number of extremely rare variants present in out-bred populations, such bottlenecked populations have greater genetic homogeneity with a simplified spectrum of rare variation: fewer total variable sites exist and those that do are more likely found in many more individuals. This observation has been borne out in numerous examples of Mendelian disease where, for example, Finns and Ashkenazi Jews have characteristic high incidence of recessive diseases because of the dramatic enrichment of specific mutations [7,8,9] – in the wider European population these same diseases are rarer and have mutational spectra involving a more diverse array of extremely rare mutations. It has not yet been assessed to which extent these population structures, so advantageous to Mendelian studies but of little importance to common variant GWAS, might generally improve the power to identify low-frequency and loss of function (LoF) variants in studies of complex disease.

To explore this question, we used exome sequencing to characterize the allelic architecture of the Finnish population compared with a set of non-Finnish Europeans (NFEs)

from the United States, Great Britain, Germany and Sweden. We demonstrate that Finns carry a significant enrichment of low-frequency (0.5-5%) loss-of-function (LoF) variation, defined here as nonsense and essential splice sites that are rare in NFEs. In addition to the isolated population structure, Finland has nationwide health records that provide decades of follow-up data that can be linked to epidemiological studies. The combination of the population structure and nationwide health records provide exceptional opportunities to study the impact of low-frequency variants on risk factors and disease outcomes and their risk factors. These opportunities have stimulated an international collaboration, The Sequencing Initiative Suomi (The SISu project) that aims to combine these resources and build knowledge and tools for genome health initiatives. We then genotyped 83 LoF variants in several large well-phenotyped population-based cohorts comprised of 36,262 Finns and tested for association to 60 quantitative traits and used data from the 13 disease outcomes assessed using the National Health Registers. We demonstrate that 5 of these variants have significant associations with clinically relevant phenotypes, demonstrating the general value of the Finnish population for the study of low-frequency variants studies in complex as well as Mendelian diseases. Using data from centralized medical records and national registries, we further confirm two LoF variants that significantly reduce Lp(a) levels are associated with protection from cardiovascular disease.

RESULTS

As part of the SISu Project, we assembled more than 5,000 whole-exome sequences from Finns in projects including GoT2D, ENGAGE, migraine, METSIM and the 1000 Genomes Project along with 3,000 whole exome-sequences of NFEs from GoT2D, ESP, NIMH and 1000 Genomes project using the same data generation and processing pipelines (Table 4.1). The raw

BAM files from these projects were compressed and re-processed at the Broad Institute and variant calling was performed in a unified manner to minimize potential batch effects. To simplify the comparison of representation of the allele frequencies, we included exactly 3,000 Finns and 3,000 NFEs in our analyses.

We initially compared the number and frequency of variable sites in 3,000 Finns and 3000 NFEs (Figure 4.1) and observed several expected hallmarks of the isolated bottlenecked Finnish population history. There was a marked depletion of 'singletons', or variants that were observed only once in 3,000 individuals, in Finns compared to NFEs – an average Finn had 3.7 times fewer singleton variants in these data (binomial $P < 1\times10^{-6}$). On the other hand, there was a marked excess of low-frequency variants in Finns versus NFEs (binomial $P < 1\times10^{-6}$), collectively suggesting that while most rare variants did not survive the bottleneck, the variants that did were of substantially elevated frequency [10], while the rates of common variation were not different between Finns and NFEs. All these findings are consistent with an expected impact of the Finnish population bottleneck.

We then stratified the variants according to their functional annotations – loss-of-function (LoF) variants, missense variants and synonymous variants. We found a higher proportion of LoF variants in Finns compared to NFEs across the rare and low-frequency allelic spectrum (Figure 4.1) and for missense variants predicted to be deleterious by PolyPhen2 (Figure 4.2). This is also a direct consequence of the bottleneck: alleles that are elevated in frequency through the bottleneck are drawn at random from extremely rare variants in the parental population, where there is a higher proportion of LoF variants that arose recently or were kept at low frequencies because of negative selection. This is clearly demonstrated with the decreasing proportions of LoF variants with increasing allele frequencies (Figure 4.1).

Table 4.1: Exomes collected from ongoing studies

Finnish studies

Project	Number of exomes
GoT2D (Botnia)	214
GoT2D (FUSION)	960
GoT2D (Scania Diabetes Registry)	9
GoT2D (Helsinki-sib)	57
GoT2D (METSIM)	962
ENGAGE (Young Finns)	717
ENGAGE (Health 2000)	
ENGAGE (FINRISK)	
1000 Genomes Project	81
Total	**3000**

NFE studies

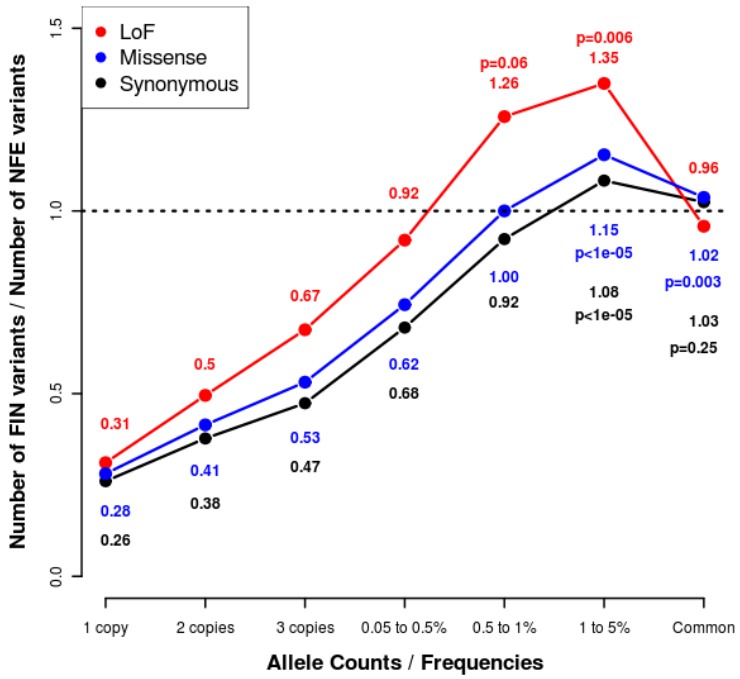
Project	Number of exomes
WTCCC	641
ESP	1792
NIMH	373
1000 Genomes Project (GBR + TSI + CEU)	194
Total	**3000**

Figure 4.1: Ratio of variants and proportion of variants found in Finns versus NFEs

(A) Ratio of the number of LoF, missense and synonymous variants found in Finns versus NFEs with the ratios for LoF variants highlighted in red text and the ratios for synonymous variants in black. The p-values represent the probabilities of the excess of variable sites in Finns occurring by chance. The p-values in red represent the probabilities for the LoF variants, the p-values in blue represent the probabilities for the missense variants and the p-values in black represent the probabilities for the synonymous variants. (B) Percentage of variants that are loss-of-function (LoF) across the allele frequency spectrum, with the numbers indicating the percentage of LoF variants in Finns versus NFEs. The p-values represent the p-values from the hypergeometric test of whether the ratio of LoF variants differ from the ratio of synonymous variants in Finns compared to NFEs.

Figure 4.1: Ratio of variants and proportion of variants found in Finns versus NFEs (Continued)

A



B

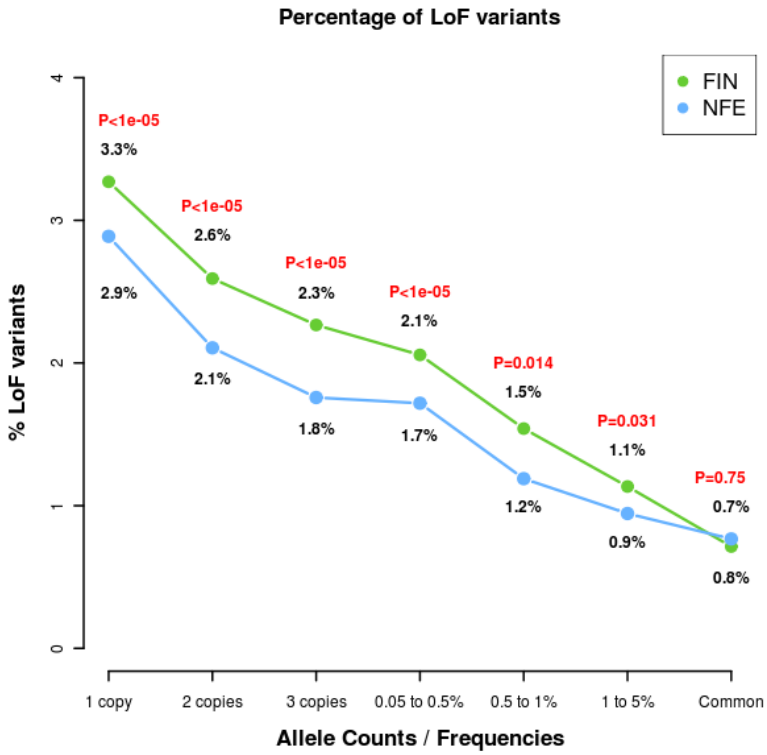
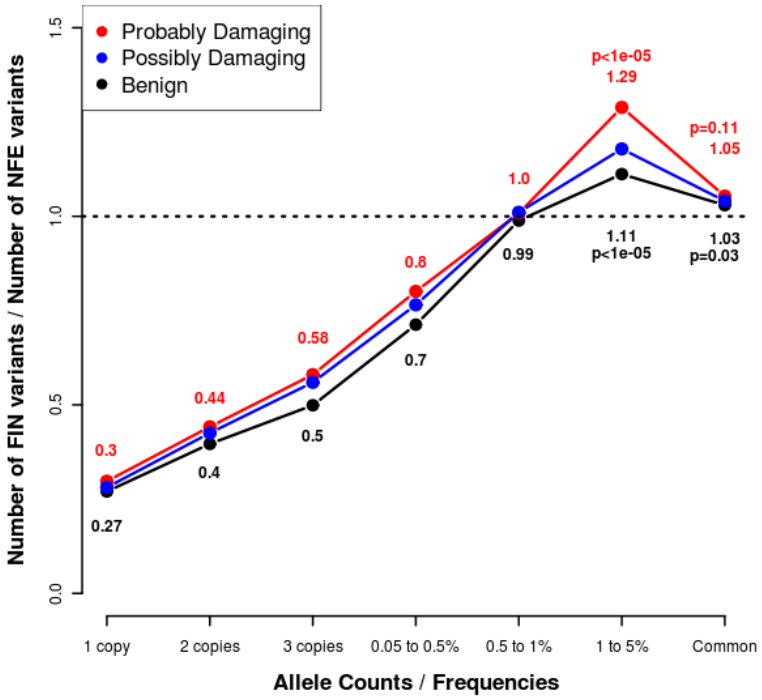


Figure 4.2: Ratio of variants and proportion of missense variants predicted by PolyPhen2 found in Finns versus NFEs

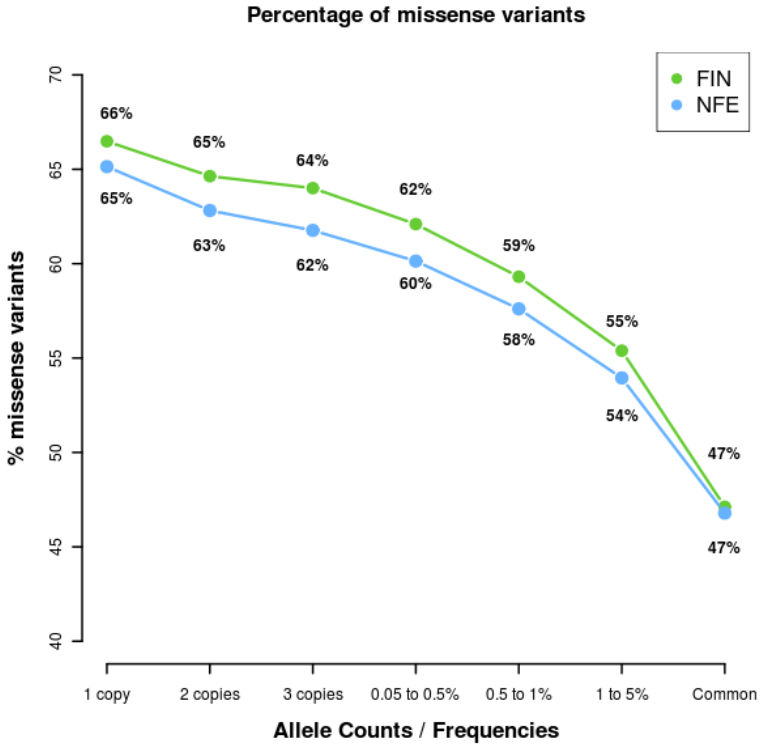
(A) The ratios for probably damaging missense variants predicted by PolyPhen2 highlighted in red text and the ratios for benign missense variants in black. The p-values represent the binomial probabilities of the variants being enriched in Finns and similarly, the p-values in red represent the probabilities for the probably damaging missense variants and the p-values in black represent the probabilities for the benign missense variants. (B) Percentage of variants that are missense variants across the allele frequency spectrum.

Figure 4.2: Ratio of variants and proportion of missense variants predicted by PolyPhen2 found in Finns versus NFEs (Continued)

A



B



The observation that LoF variants in the 0.5-5% range are enriched in Finns and our hypothesis that some of these variants might have strong phenotypic consequences, motivated the large targeted association study described below (Figure 4.3). Despite the reduced overall variation in the isolated population, the existence of a greater number of low frequency LoF variants results in an average Finn harboring 0.16 homozygous LoF variants compared to only 0.095 in an average NFE, driven primarily by homozygosity in the 0.5 to 5% allele frequency range (Figure 4.4). These features of the Finnish population have already been well described as they pertain to Mendelian diseases: many characteristic "Finnish founder mutations" exist at unusually high frequencies, even up to 1%, for highly penetrant and reproductively lethal disorders while such variants are extremely rare or absent in NFEs [11]. We confirmed with simulations that while such variants are inevitably pushed to extremely low frequency after 1,000 or more generations, they can easily persist at frequencies between 0.1 and 1% up to 100 generations after a bottleneck (Figure 4.5). Table 4.2 shows a table of a set of Finnish Disease Heritage (www.findis.org) variants and their population frequencies. The extent to which such variants contribute to more common diseases, either through highly-penetrant recessive subtypes or modest risk to carriers, will correspond to advantages in rare and low-frequency association studies in isolated populations.

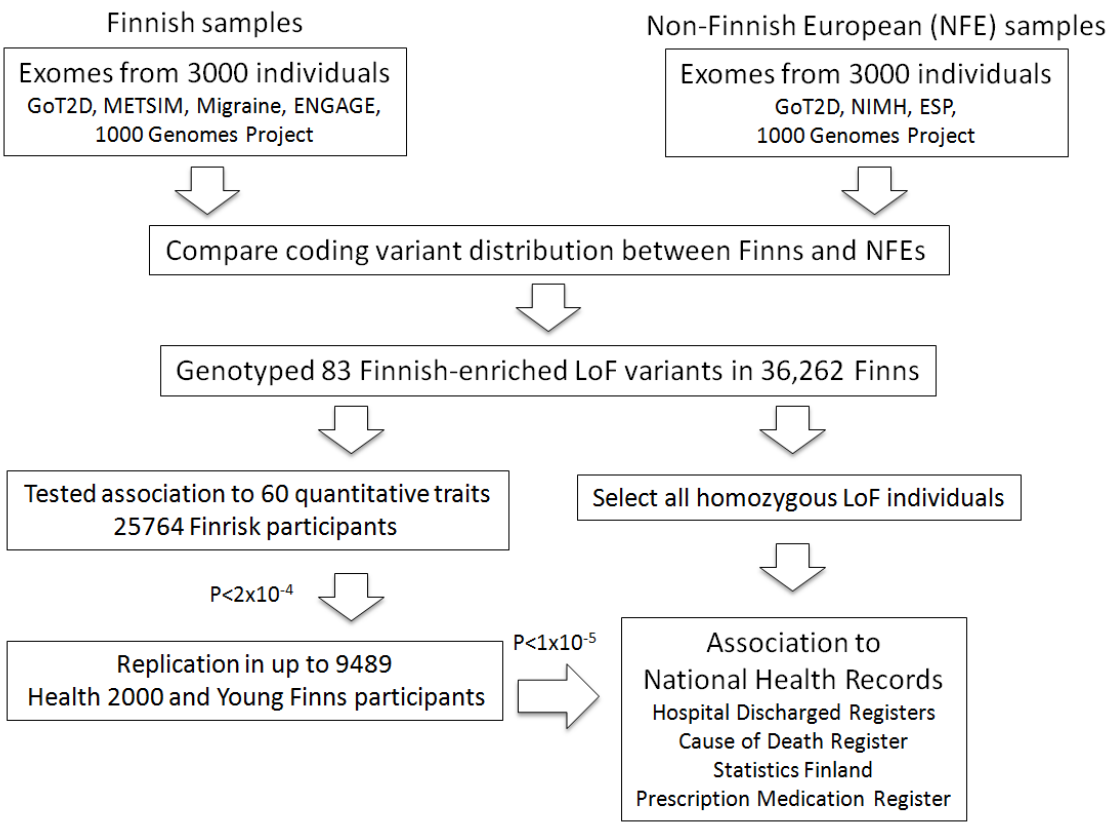
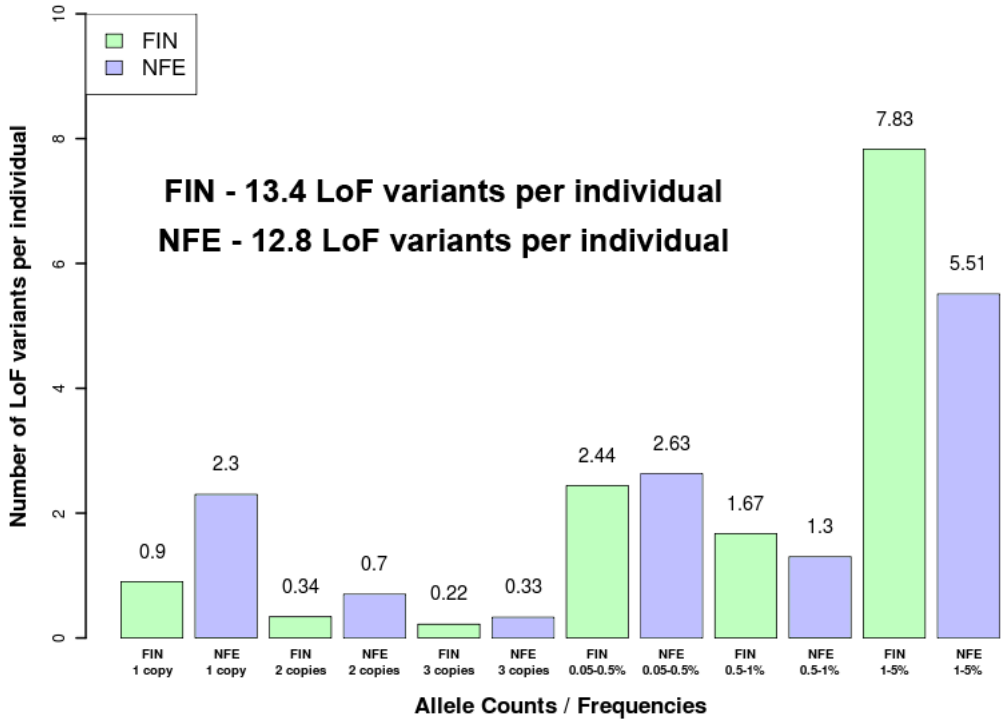


Figure 4.3: Study design figure illustrating the analysis

We used an initial set of exome sequences from Finns and NFEs to perform the selection and survey of the 83 LoF variants across 60 quantitative traits and 13 disease categories.

A



B

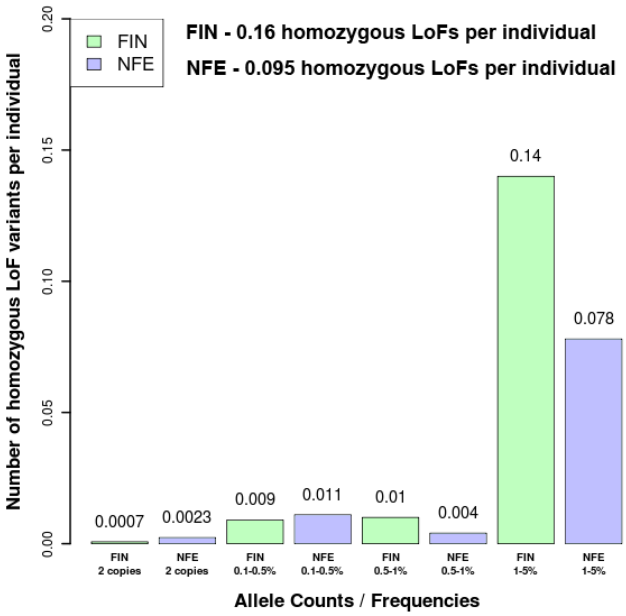


Figure 4.4: Number of LoF variants per individual

(A) Number of LoF variants in Finns vs NFEs per individual. (B) Number of homozygous LoF variants in Finns vs NFEs per individual.

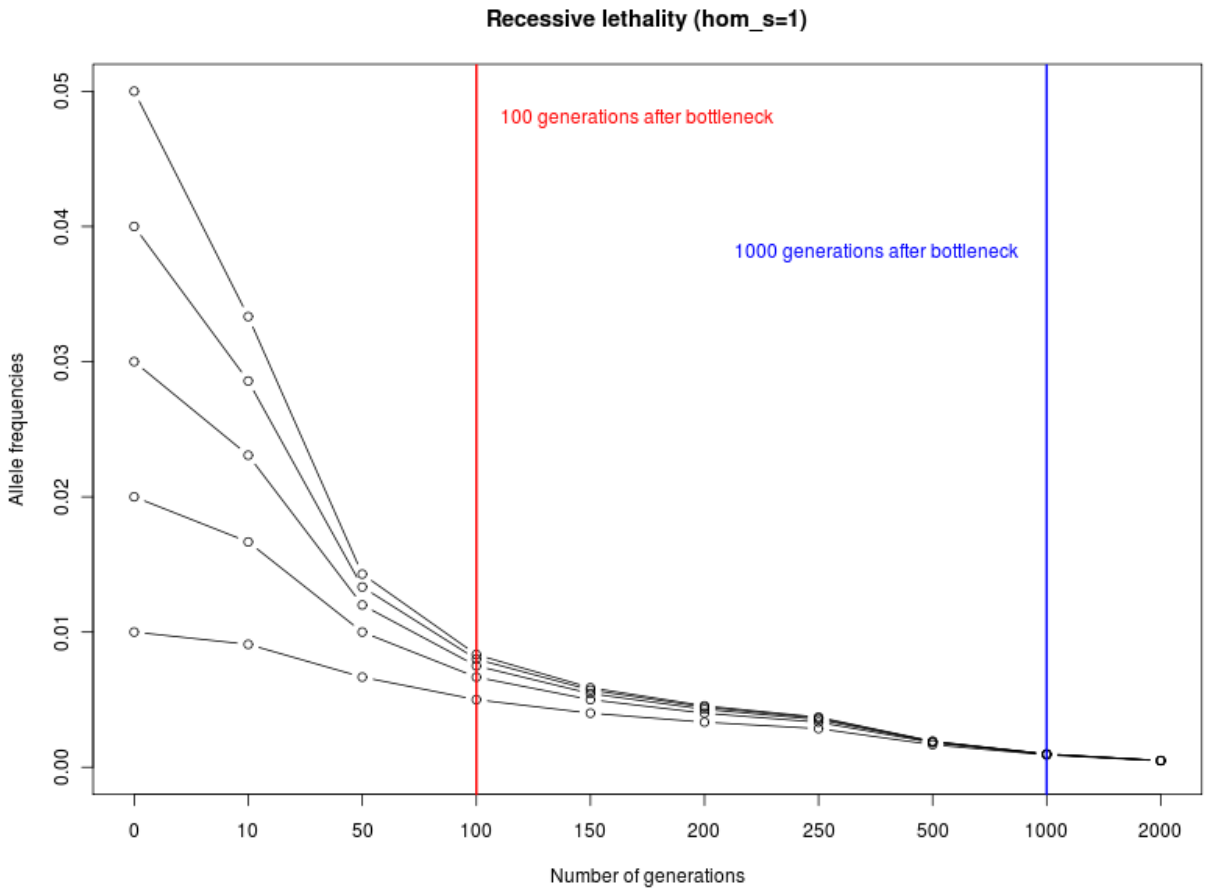


Figure 4.5: Simulations for a set of variants (ranging from 1% to 5% allele frequencies) with complete recessive lethality

The red line indicates the expected allele frequencies in present-day Finns (where the Finnish bottleneck occurred ~100 generations ago) and the blue line indicates the expected allele frequencies in Finns 1,000 generations after the Finnish bottleneck, similar to the out-of-Africa bottleneck which occurred >1,000 generations ago.

Table 4.2: Allele frequencies of variants discovered from the FinDis database

Chr	Pos	Gene	Type	AA	AA_Pos	FInn_AF	NFE_AF	Disease
9	6588403	GLDC	NON_SYNONYMOUS_CODON	A/T	569	0.02251	0.00583	Glycine encephalopathy
9	6556242	GLDC	NON_SYNONYMOUS_CODON	V/M	705	0.01468	0.0025	Glycine encephalopathy
11	125769895	HYLS1	NON_SYNONYMOUS_CODON	D/G	211	0.00884	0.00017	Hydrolethalus syndrome 1
9	131284937	GLE1	INTRON_VARIANT	-	-	0.00855	0	Lethal congenital contracture syndrome 1;Arthrogryposis, lethal, with anterior horn cell disease
2	49210264	FSHR	NON_SYNONYMOUS_CODON	A/V	189	0.00785	0	Ovarian dysgenesis 1
13	77574606	CLN5	NON_SYNONYMOUS_CODON	N/K	108	0.00784	0.01568	Ceroid lipofuscinosis, neuronal, 5
1	40557070	PPT1	NON_SYNONYMOUS_CODON	R/W	122	0.00736	0	Ceroid lipofuscinosis, neuronal, 1
2	136564701	LCT	STOP_GAINED	Y/*	822	0.00717	0	Lactase deficiency, congenital
17	57157240	TRIM37	SPLICE_ACCEPTOR_VARIANT	-	-	0.007	0	Mulibrey nanism
3	150645894	CLRN1	STOP_GAINED	Y/*	176	0.00684	0	Usher syndrome, type 3A
10	17113456	CUBN	NON_SYNONYMOUS_CODON	S/N	865	0.00652	0.01084	Megaloblastic anemia-1, Finnish type
4	178359918	AGA	NON_SYNONYMOUS_CODON	C/S	163	0.00567	0	Aspartylglucosaminuria
4	178359924	AGA	NON_SYNONYMOUS_CODON	R/Q	161	0.00567	0	Aspartylglucosaminuria
6	74354306	SLC17A5	NON_SYNONYMOUS_CODON	R/C	39	0.00517	0.00067	Sialuria, Finnish type (Salla disease)
13	77566147	CLN5	NON_SYNONYMOUS_CODON	P/S	21	0.00412	0.00505	Ceroid lipofuscinosis, neuronal, 5

Table 4.2: Allele frequencies of variants discovered from the FinDis database (Continued)

Chr	Pos	Gene	Type	AA	AA_Pos	FInn_AF	NFE_AF	Disease
10	17083159	CUBN	NON_SYNONYMOUS_CODON	P/L	1297	0.004	0	Megaloblastic anemia-1, Finnish type
14	23245147	SLC7A7	SPLICE_ACCEPTOR_VARIANT	-	-	0.004	0	Lysinuric protein intolerance
2	219525942	BCS1L	NON_SYNONYMOUS_CODON	S/G	78	0.00383	0.00017	GRACILE syndrome
21	45709656	AIRE	STOP_GAINED	R/*	257	0.00353	0.00051	Autoimmune polyendocrinopathy syndrome , type I, with or without reversible metaphyseal dysplasia
12	91449319	KERA	NON_SYNONYMOUS_CODON	N/S	247	0.0032	0	Cornea plana 2
10	126086626	OAT	NON_SYNONYMOUS_CODON	L/P	264	0.003	0	Gyrate atrophy of choroid and retina with or without ornithinemia
10	126086626	OAT	NON_SYNONYMOUS_CODON	L/P	264	0.003	0	Gyrate atrophy of choroid and retina with or without ornithinemia
4	15538697	CC2D2A	STOP_GAINED	Q/*	31	0.00255	0	Meckel syndrome 6
1	46657769	POMGNT1	SPLICE_DONOR_VARIANT	-	-	0.00234	0.00017	Muscular dystrophy-dystroglycanopathy (congenital with brain and eye anomalies), type A, 3
1	46657769	POMGNT1	SPLICE_DONOR_VARIANT	-	-	0.00234	0.00017	Muscular dystrophy-dystroglycanopathy (congenital with brain and eye anomalies), type A, 3

Table 4.2: Allele frequencies of variants discovered from the FinDis database (Continued)

Chr	Pos	Gene	Type	AA	AA_Pos	FInn_AF	NFE_AF	Disease
10	102750231	C10orf2	NON_SYNONYMOUS_CODON	Y/C	508	0.00183	0	Mitochondrial DNA depletion syndrome 7 (hepatocerebral type)
17	56292121	MKS1	NON_SYNONYMOUS_CODON	R/W	156	0.00137	0	Meckel syndrome 1
8	100832259	VPS13B	NON_SYNONYMOUS_CODON	N/S	2993	0.00117	0.00717	Cohen syndrome
16	53639522	RPGRIP1L	NON_SYNONYMOUS_CODON	R/C	1202	0.001	0.00067	Meckel syndrome 5
1	40557075	PPT1	SPLICE_REGION_VARIANT	-	-	0.00084	0.00134	Ceroid lipofuscinosis, neuronal, 1
9	6554703	GLDC	NON_SYNONYMOUS_CODON	G/R	761	0.00084	0	Glycine encephalopathy
9	6588417	GLDC	NON_SYNONYMOUS_CODON	S/I	564	0.00083	0	Glycine encephalopathy
5	149360691	SLC26A2	NON_SYNONYMOUS_CODON	T/K	512	0.00067	0	Diastrophic dysplasia
10	126094114	OAT	NON_SYNONYMOUS_CODON	R/T	42	0.00067	0	Gyrate atrophy of choroid and retina with or without ornithinemia
19	36322260	NPHS1	STOP_GAINED	R/*	1109	0.00067	0	Nephrotic syndrome, type 1
9	131303403	GLE1	NON_SYNONYMOUS_CODON	I/T	684	0.0005	0	Arthrogryposis, lethal, with anterior horn cell disease
1	46655129	POMGNT1	SPLICE_DONOR_VARIANT	-	-	0.00034	0.00017	Muscular dystrophy-dystroglycanopathy (congenital with brain and eye anomalies), type A, 3
1	46655129	POMGNT1	SPLICE_DONOR_VARIANT	-	-	0.00034	0.00017	Muscular dystrophy-dystroglycanopathy

Table 4.2: Allele frequencies of variants discovered from the FinDis database (Continued)

Chr	Pos	Gene	Type	AA	AA_Pos	FInn_AF	NFE_AF	Disease
								(congenital with brain and eye anomalies), type A, 3
5	149359991	SLC26A2	NON_SYNONYMOUS_CODON	R/W	279	0.00033	0.00133	Diastrophic dysplasia
5	149357444	SLC26A2	NON_SYNONYMOUS_CODON	N/H	77	0.00017	0.0005	Diastrophic dysplasia
7	107423465	SLC26A3	NON_SYNONYMOUS_CODON	S/F	398	0.00017	0	Chloride diarrhea, congenital, Finnish type
8	94808198	TMEM67	NON_SYNONYMOUS_CODON	C/R	534	0.00017	0	Meckel syndrome 3
9	6605186	GLDC	NON_SYNONYMOUS_CODON	T/M	269	0.00017	0	Glycine encephalopathy
9	131298693	GLE1	NON_SYNONYMOUS_CODON	R/H	569	0.00017	0.00033	Lethal congenital contracture syndrome 1
12	91445245	KERA	STOP_GAINED	R/*	313	0.00017	0	Cornea plana 2
14	23282447	SLC7A7	NON_SYNONYMOUS_CODON	G/V	54	0.00017	0	Lysinuric protein intolerance
16	28493481	CLN3	NON_SYNONYMOUS_CODON	R/H	234	0.00017	0	Ceroid lipofuscinosis, neuronal, 3
1	40555089	PPT1	NON_SYNONYMOUS_CODON	Q/E	177	0	0.00017	Ceroid lipofuscinosis, neuronal, 1
1	40555167	PPT1	STOP_GAINED	R/*	151	0	0.00067	Ceroid lipofuscinosis, neuronal, 1
1	40562882	PPT1	STOP_GAINED	L/*	10	0	0.00017	Ceroid lipofuscinosis, neuronal, 1
1	46658068	POMGNT1	NON_SYNONYMOUS_CODON	R/H	442	0	0.00033	Muscular dystrophy-dystroglycanopathy (congenital with brain and eye anomalies), type A, 3
2	136558283	LCT	NON_SYNONYMOUS_CODON	R/H	1587	0	0.00017	Lactase deficiency, congenital

Table 4.2: Allele frequencies of variants discovered from the FinDis database (Continued)

Chr	Pos	Gene	Type	AA	AA_Pos	FInn_AF	NFE_AF	Disease
2	219525876	BCS1L	STOP_GAINED	R/*	56	0	0.00017	GRACILE syndrome
3	150690352	CLRN1	NON_SYNONYMOUS_CODON	N/K	48	0	0.00017	Usher syndrome, type 3A
4	15572069	CC2D2A	NON_SYNONYMOUS_CODON	W/R	625	0	0.00018	Meckel syndrome 6
5	149357747	SLC26A2	STOP_GAINED	R/*	69	0	0.00033	Diastrophic dysplasia
5	149361150	SLC26A2	NON_SYNONYMOUS_CODON	H/P	665	0	0.00017	Diastrophic dysplasia
8	1719266	CLN8	NON_SYNONYMOUS_CODON	L/M	16	0	0.00017	Ceroid lipofuscinosis, neuronal, 8
8	1719594	CLN8	NON_SYNONYMOUS_CODON	N/S	125	0	0.00133	Ceroid lipofuscinosis, neuronal, 8
8	1728557	CLN8	NON_SYNONYMOUS_CODON	P/A	229	0	0.00017	Ceroid lipofuscinosis, neuronal, 8
8	94777876	TMEM67	SPLICE_DONOR_VARIANT	-	-	0	0.00017	Meckel syndrome 3
8	94817024	TMEM67	NON_SYNONYMOUS_CODON	G/E	705	0	0.00017	Meckel syndrome 3
8	100155318	VPS13B	NON_SYNONYMOUS_CODON	A/T	590	0	0.00183	Cohen syndrome
8	100729602	VPS13B	SPLICE_DONOR_VARIANT	-	-	0	0.00033	Cohen syndrome
8	100830757	VPS13B	STOP_GAINED	R/*	2839	0	0.00017	Cohen syndrome
9	6553420	GLDC	NON_SYNONYMOUS_CODON	A/V	802	0	0.00017	Glycine encephalopathy
9	6587141	GLDC	NON_SYNONYMOUS_CODON	S/T	617	0	0.00017	Glycine encephalopathy
9	6595109	GLDC	NON_SYNONYMOUS_CODON	A/V	389	0	0.00018	Glycine encephalopathy
9	6602147	GLDC	NON_SYNONYMOUS_CODON	R/W	373	0	0.00017	Glycine encephalopathy
10	17152923	CUBN	NON_SYNONYMOUS_CODON	P/L	337	0	0.00017	Megaloblastic anemia-1, Finnish type

Table 4.2: Allele frequencies of variants discovered from the FinDis database (Continued)

Chr	Pos	Gene	Type	AA	AA_Pos	FInn_AF	NFE_AF	Disease
10	102749544	C10orf2	NON_SYNONYMOUS_CODON	R/W	463	0	0.00017	Mitochondrial DNA depletion syndrome 7 (hepatocerebral type)
10	126086581	OAT	NON_SYNONYMOUS_CODON	P/L	279	0	0.00017	Gyrate atrophy of choroid and retina with or without ornithinemia
10	126086581	OAT	NON_SYNONYMOUS_CODON	P/L	279	0	0.00017	Gyrate atrophy of choroid and retina with or without ornithinemia
10	126092416	OAT	NON_SYNONYMOUS_CODON	P/L	103	0	0.00033	Gyrate atrophy of choroid and retina with or without ornithinemia
10	126100579	OAT	NON_SYNONYMOUS_CODON	N/K	54	0	0.00017	Gyrate atrophy of choroid and retina with or without ornithinemia
13	77566309	CLN5	NON_SYNONYMOUS_CODON	W/R	75	0	0.00144	Ceroid lipofuscinosis, neuronal, 5
13	77570169	CLN5	NON_SYNONYMOUS_CODON	W/R	73	0	0.00017	Ceroid lipofuscinosis, neuronal, 5
13	77570221	CLN5	STOP_GAINED	W/*	90	0	0.00017	Ceroid lipofuscinosis, neuronal, 5
16	53679606	RPGRIP1L	STOP_GAINED	Q/*	872	0	0.00017	Meckel syndrome 5
17	19251095	B9D1	SPLICE_DONOR_VARIANT	-	-	0	0.00017	Meckel syndrome 9
17	56293449	MKS1	SPLICE_REGION_VARIANT	E	129	0	0.00017	Meckel syndrome 1
21	45194641	CSTB	SPLICE_ACCEPTOR_VARIANT	-	-	0	0.00033	Epilepsy, progressive myoclonic 1A (Unverricht and Lundborg)

Given our empirical observations of proportionally more LoF variants in the 0.5-5% allele frequency range in Finns, we next conducted a pilot test of this hypothesis that some of the Finnish-enriched low-frequency LoF variants might have strong phenotypic effects. We successfully genotyped 83 LoF variants (protein-truncating nonsense, essential splice site variants and frameshift variants) that were enriched in Finns. Of these 83 variants, 76 variants were more than 2-fold enriched and 26 were more than 10-fold enriched, and 75 out of the 83 LoF variants found at 0.5-5% allele frequency in Finns) using Sequenom MALDI-TOF genotyping assays (Table 4.3). Three genes (*SERPINA10*, *LPA* and *FANCM*) contained two LoF variants each; we combined these pairs and tested them as single composite LoF variants, resulting in a total of 80 independent LoF variants tested in this study. These 83 variants were genotyped in a total of 36.262 individuals from three population cohorts: FINRISK [12] (26,245 individuals), Health2000 (7,363 individuals) and Young Finns [13] (2,654 individuals).

As these three studies are population-based cohorts, we were able to assess whether any of the homozygous LoF variants might be lethal in fetal life or early infancy, or result in such a severe phenotype that these individuals would not be able to participate in a population survey. Study-wide, there was a modest excess of homozygotes of the variants (1.23-fold versus Hardy-Weinberg expectation) arising from within population substructure. A nonsense variant in the *Translation Elongation Factor, Mitochondrial gene* (*TSFM*) that is present at 1.2% allele frequency in Finns and absent in NFEs, was not found in a homozygous state in >36,000 Finns (Hardy Weinberg Equilibrium (HWE) P = 0.0077). This suggests that complete loss of *TSFM* might result in embryonic lethality, severe childhood diseases in humans, or that the individuals might not have been ascertained by the studies employed, i.e. if the individuals are too sick to be included in the studies. A lookup of this variant in another 25,237 Finnish exome chip

genotyping data from the GoT2D studies confirmed that the variant is present at 1.2% in Finns, but again with no homozygotes observed (combined HWE P = 1.6×10^{-4}). The fact that recessive missense variants in *TSFM* have been reported to result in mitochondrial translation deficiency [14,15] lends additional plausibility to the hypothesis that complete loss of this gene is not tolerated in humans.

Several other LoF variants occur in genes where recessive mutations have been noted to cause severe Mendelian diseases from the Online Mendelian Inheritance in Man database (OMIM) [16]. For instance, the *Fanconi anemia complementation group M* gene (*FANCM*) was initially discovered in one family with Fanconi anemia [17], but we did not observe any deficit of homozygous LoFs in *FANCM* from our dataset (expected = 5, observed = 7), which we would typically observe for a disease causing recessive variant. However, examination of the hospital discharge records did not provide any evidence for blood diseases, increased cancer events or any other chronic diseases in these individuals with homozygous LoFs in *FANCM*. Singh *et al.* reported that the initial case that led to the association of *FANCM* with Fanconi anemia also harbor biallelic, functional mutations in *FANCA*, a well established Fanconi anemia gene [18]. Our findings in this study, combined with the findings by Singh *et al.* do not support the hypothesis that *FANCM* is a Fanconi anemia gene.

Table 4.3: Final list of variants from Sequenom genotyping in 36,262 Finns

Chr	Type	Gene	Annotation	#Hom ref	#Het	#Hom nonref	Genotyping Finn_AF	Sequencing Finn_AF	Sequencing NFE_AF	Expected Het	Expected Hom	HWE
1	frameshift	PADI1	-	35238	519	3	0.00734	0.01131	9.00E-04	521	2	0.4465
1	stop	COL9A2	full	34981	798	4	0.01126	0.01105	0.00316	797	5	1.0000
1	stop	SLFNL1	full	33429	1617	29	0.02388	0.03517	0.00676	1635	20	0.0500
1	frameshift	GBP5	-	34496	641	4	0.00923	0.01010	0.00224	643	3	0.5485
1	splice	DPYD	partial	33517	1465	20	0.02150	0.01659	0.00268	1473	16	0.3088
1	splice	AKNAD1	partial	33619	2124	36	0.03069	0.03762	0.00045	2129	34	0.6565
1	frameshift	C1orf56	-	33788	707	7	0.01045	0.01252	0.00179	713	4	0.1075
1	stop	SELP	partial	34926	866	9	0.01235	0.01134	0	873	5	0.1253
1	stop	CFHR2	full	33171	1820	31	0.02687	0.03519	0.00671	1831	25	0.2592
1	stop	OBSCN	partial	33419	2341	44	0.03392	0.02743	0.00305	2347	41	0.6282
1	frameshift	LGALS8	-	33104	1959	36	0.02893	0.03514	0	1972	29	0.2155
2	stop	GCA	partial	34908	893	11	0.01278	0.01598	0.00179	903	6	0.0532
2	frameshift	PDE11A	-	34222	898	7	0.01298	0.01616	0.00313	900	6	0.5351
2	frameshift	DNAH7	-	34936	801	8	0.01143	0.01051	0.00134	808	5	0.1498
2	stop	HTR2B	partial	34613	1181	9	0.01674	0.01172	0	1179	10	0.8728
3	stop	CCDC37	partial	33549	2171	48	0.03169	0.02991	0.00514	2195	36	0.0467

Table 4.3: Final list of variants from Sequenom genotyping in 36,262 Finns (Continued)

Chr	Type	Gene	Annotation	#Hom ref	#Het	#Hom nonref	Genotyping Finn_AF	Sequencing Finn_AF	Sequencing NFE_AF	Expected Het	Expected Hom	HWE
3	stop	A4GNT	full	34472	569	4	0.00823	0.01132	0.00089	572	2	0.3057
4	splice	CDKL2	full	34251	1490	23	0.02147	0.01948	0.00492	1503	16	0.1015
4	stop	HERC6	partial	33345	1733	26	0.02542	0.02061	0.00089	1740	23	0.4503
5	frameshift	PCDHA3	-	33410	2344	47	0.03405	0.03110	0.00671	2355	42	0.3760
5	frameshift	FCHSD1	-	34150	941	14	0.01380	0.01494	0.00179	956	7	0.0091
5	stop	GPR151	full	34502	1256	15	0.01797	0.01940	0.00537	1263	12	0.2919
5	splice	ARHGEF37	full	34051	1056	10	0.01532	0.01617	0.0076	1060	8	0.4779
5	stop	ATP10B	full	32747	2292	46	0.03397	0.02932	0.00045	2303	40	0.3701
6	frameshift	GPLD1	-	34388	1415	19	0.02028	0.01737	0.00492	1424	15	0.2322
6	stop	SLC17A4	partial	34014	1772	22	0.02536	0.02510	0.01565	1770	23	0.9150
6	stop	CRISP1	full	34293	855	8	0.01239	0.01262	0.00134	860	5	0.2673
6	frameshift	MYCT1	-	31828	3825	112	0.05661	0.06093	0.02507	3820	115	0.8430
6	stop	CLDN20	full	33768	2049	37	0.02961	0.02710	0	2060	31	0.3108
6	splice	LPA	full	33712	1883	45	0.02768	0.02183	0.00492	1918	27	0.0011
6	splice	LPA	full	32886	3294	82	0.04768	0.04689	0.03533	3293	82	1.0000
6	-	LPA	-	31067	4968	227	0.07476	-	-	5017	203	0.0683

Table 4.3: Final list of variants from Sequenom genotyping in 36,262 Finns (Continued)

Chr	Type	Gene	Annotation	#Hom ref	#Het	#Hom nonref	Genotyping Finn_AF	Sequencing Finn_AF	Sequencing NFE_AF	Expected Het	Expected Hom	HWE
6	splice	WDR27	partial	34109	938	6	0.01355	0.01217	0.00586	937	6	1.0000
7	frameshift	TMEM195	-	34659	1114	17	0.01604	0.02102	0.00268	1130	9	0.0169
7	frameshift	PRPS1L1	-	32248	3386	101	0.05020	0.05335	0.02415	3408	90	0.2222
7	splice	ABCB5	partial	30082	5388	267	0.08286	0.06178	0.0083	5431	245	0.1344
7	stop	POM121L12	full	32808	2527	78	0.03788	0.02660	9.00E-04	2581	51	0.0002
7	stop	CCL26	full	34930	891	6	0.01260	0.02425	0.01342	892	6	0.8302
7	stop	C7orf64	partial	30517	4361	137	0.06619	0.06854	0.02816	4328	153	0.1664
7	stop	NDUFA5	partial	31994	2997	68	0.04468	0.04082	0.00568	2993	70	0.8512
7	stop	CLCN1	full	34632	1120	13	0.01602	0.01536	0.00045	1128	9	0.1805
8	frameshift	FGL1	-	32954	2514	63	0.03715	0.03877	0.00045	2542	49	0.0447
8	frameshift	HTRA4	-	34843	930	10	0.01327	0.01373	0.00313	937	6	0.1510
8	frameshift	EPPK1	-	34154	1619	21	0.02320	0.02060	0	1622	19	0.6406
9	stop	IFNA5	full	34211	1543	23	0.02221	0.01657	0.00581	1554	18	0.1806
9	stop	IFNE	full	33924	1134	14	0.01657	0.01904	0.00134	1143	10	0.1408
9	splice	SOHLH1	full	34816	992	15	0.01426	0.01173	0	1007	7	0.0078
9	splice	PNPLA7	partial	34442	1333	16	0.01907	0.01909	0.00762	1339	13	0.3938

Table 4.3: Final list of variants from Sequenom genotyping in 36,262 Finns (Continued)

Chr	Type	Gene	Annotation	#Hom ref	#Het	#Hom nonref	Genotyping Finn_AF	Sequencing Finn_AF	Sequencing NFE_AF	Expected Het	Expected Hom	HWE
10	frameshift	PTCHD3	-	32838	2182	52	0.03259	0.04058	0.00493	2211	37	0.0173
10	stop	ZMYND17	full	34967	817	7	0.01161	0.01819	0.00179	821	5	0.3475
10	stop	LIPK	full	34437	673	6	0.00975	0.01011	0.00224	678	3	0.1539
10	frameshift	CALHM2	-	33416	2293	48	0.03341	0.03796	0.00939	2309	40	0.1893
11	splice	MS4A2	full	33769	1307	15	0.01905	0.01377	0.00089	1312	13	0.4743
11	stop	P4HA3	partial	16564	15276	3795	0.32084	0.31727	0.15229	15530	3668	0.0021
12	splice	PRPH	full	34170	1553	23	0.02237	0.01617	0.00539	1563	18	0.2242
12	frameshift	SPATS2	-	34419	726	7	0.01053	0.01212	0.00134	732	4	0.1169
12	stop	TSFM	partial	34928	886	0	0.01237	0.01184	0	875	5	0.0077
13	stop	ZMYM5	full	34287	809	9	0.01178	0.01375	0.00134	817	5	0.0647
13	stop	CLYBL	full	30879	2210	39	0.03453	0.03759	0.03491	2209	40	1.0000
14	stop	FANCM	full	35122	636	5	0.00903	0.01293	0	640	3	0.2216
14	stop	FANCM	full	36060	201	1	0.00280	0.00405	0.00089	202	0	0.2469
14	-	FANCM	-	35422	833	7	0.01168	-	-	837	5	0.3541
14	stop	TBPL2	full	34176	780	11	0.01147	0.01091	0	793	5	0.0070
14	stop	SERPINA10	full	34505	1199	11	0.01709	0.01659	0.00492	1200	10	0.7537

Table 4.3: Final list of variants from Sequenom genotyping in 36,262 Finns (Continued)

Chr	Type	Gene	Annotation	#Hom ref	#Het	#Hom nonref	Genotyping Finn_AF	Sequencing Finn_AF	Sequencing NFE_AF	Expected Het	Expected Hom	HWE
14	stop	SERPINA10	full	36170	92	0	0.00127	0.00081	0.0076	92	0	1.0000
14	-	SERPINA10	-	34965	1283	14	0.01808	-	-	1287	12	0.4604
14	stop	SERPINA12	full	34284	817	6	0.01181	0.01657	0.00716	819	5	0.4955
15	frameshift	PLCB2	-	34675	1014	8	0.01443	0.01292	0	1015	7	0.7110
15	frameshift	DUOX2	-	34415	664	6	0.00963	0.01334	0.00269	669	3	0.1469
15	stop	GCNT3	full	34994	744	9	0.01066	0.01496	0.00089	754	4	0.0218
16	splice	CCDC78	partial	35046	633	6	0.00904	0.01536	0.00089	639	3	0.0733
16	stop	PHKB	partial	34739	1019	15	0.01466	0.01444	0.00048	1034	8	0.0150
16	stop	ELMO3	full	34047	1692	26	0.02438	0.01657	0.00268	1701	21	0.2670
16	splice	ATP2C2	full	34178	1563	28	0.02263	0.02061	0.0085	1582	18	0.0301
17	splice	EFCAB5	partial	34551	1217	13	0.01737	0.01602	0.00583	1221	11	0.4388
17	stop	EFCAB3	full	31576	2831	63	0.04289	0.03880	0.00626	2830	63	1.0000
18	frameshift	MRO	-	33955	1154	19	0.01697	0.02342	0.00984	1172	10	0.0092
19	stop	ZNF763	full	33920	1073	14	0.01573	0.01334	0.00134	1084	9	0.0801
19	stop	ZNF571	full	32746	3003	77	0.04406	0.03930	0.00403	3018	70	0.3465
19	splice	SPHK2	partial	34249	1418	14	0.02026	0.01870	0.00328	1417	15	1.0000

Table 4.3: Final list of variants from Sequenom genotyping in 36,262 Finns (Continued)

Chr	Type	Gene	Annotation	#Hom ref	#Het	#Hom nonref	Genotyping Finn_AF	Sequencing Finn_AF	Sequencing NFE_AF	Expected Het	Expected Hom	HWE
19	stop	FUT2	full	3393	12254	9256	0.61772	0.37662	0.47139	11761	9502	3.69E-11
19	frameshift	NLRP13	-	35129	679	3	0.00956	0.01131	0.00179	678	3	1.0000
19	stop	ZNF772	partial	33489	1613	22	0.02359	0.01739	0.00894	1618	20	0.5610
19	stop	ZNF544	partial	32421	2632	69	0.03943	0.03597	0.01118	2661	55	0.0482
20	frameshift	FAM65C	-	34242	1510	18	0.02161	0.01817	0	1513	17	0.7069
22	frameshift	DNAJB7	-	33281	1758	29	0.02589	0.01939	0.00492	1769	24	0.2430

Mutations in *COL9A2* can cause autosomal dominant multiple epiphyseal dysplasia or autosomal recessive Stickler syndrome, diseases that affect the connective tissues and result in underdeveloped bones, among several other phenotypes. When surveying data from two major RNA sequencing studies [19,20], we observed the Q326X variant in the Collagen, Type IX, Alpha 2 gene (*COL9A2*) to be associated to decreased RNA levels in several tissues, including lung, liver and skin, from the RNA sequencing data (binomial $P = 3.51\times10^{-9}$). However, we did not detect any deviation from HWE for the *COL9A2* Q326X variant and the hospital discharge records did not provide any evidence for connective tissue diseases with the 798 heterozygous carriers or the 4 homozygous carriers. Likewise, *Dihydropyrimidine dehydrogenase* (*DPYD*) deficiency causes intolerance to overload of pyrimidines, such as the chemotherapeutic agent 5-fluorouracil [21]. Some reports suggest that *DPYD* deficient homozygotes might be associated with neurological abnormalities [22], whereas others do not find any specific phenotype associations [23]. In our study, the hospital discharge records did not indicate any enrichment of any disease categories among the 20 individuals with homozygous for LoFs in *DPYD*.

The FINRISK cohort had collected 60 biochemical and physiological quantitative measurements of cardiovascular or immunologic relevance (Table 4.4), some of which are highly correlated. We tested the 80 variants across the 60 traits and report from this initial screen all associations with $p<2\times10^{-4}$ – that is, a value where we would expect only one chance observation in the entire study. In total, we observed 41 associations that exceeded this significance threshold (Table 4.5), far beyond the expected. If the phenotype was available in the Young Finns and Health 2000 cohorts, replication was attempted for these initial scan hits and significant associations are highlighted below when the combined p-value was smaller than a conservative study-wide Bonferroni-corrected threshold of $0.05/(80*60) = 1\times10^{-5}$.

Table 4.4: List of 60 blood pressure measures and biochemical assays

Trait	Description
Systolic bp	Systolic blood pressure
Diastolic bp	Diastolic blood pressure
HDL	High density lipoprotein
Triglycerides	Triglycerides
LDL	Low density lipoprotein
Lp(a)	Lipoprotein (a)
APOA1	Apolipoprotein A-I
APOB	Apolipoprotein B
Galectin-3	Galectin-3
LPS	Lipopolysaccharide
CRP	C reactive protein
HGF	Hepatocyte growth factor
SCF	Stem cell factor
SDF1	Stromal cell derived factor 1 (CXCL12)
TNF-beta	Tumor necrosis factor beta
TRAIL	TNF related apoptosis inducing ligand
IL4	Interleukin 4
IL6	Interleukin 6
IL10	Interleukin 10
IL12	Interleukin 12
IL17	Interleukin 17
Eotaxin	Eotaxin (CCLL11)

Table 4.4: List of 60 blood pressure measures and biochemical assays (Continued)

Trait	Description
FGF	Fibroblast growth factor
GCSF	Granulocyte colony stimulating factor
GM-CSF	Granulocyte monocyte colony stimulating factor
IFN-gamma	Interferon_gamma
MCP1	Monocyte chemoattractant protein 1 (CCL2)
PDGF	Platelet derived growth factor
MIP1B	Macrophage inflammatory protein 1 beta (CCL4)
RANTES	Regulated on Activation, Normal T cell Expressed and Secreted (CCL5)
VEGF	Vascular endothelial cell growth factor A
Active-B12	Active vitamin B12
Adiponectin	Adiponectin
BNP	Brain_natriuretic_peptide
CK_MB	Creatine kinase isoenzyme MB
Creatinine	Creatinine
Vasopressin	C terminal pro vasopressin
Endothelin1	C terminal pro endothelin 1
Cystatin-C	Cystatin C
D-dimer	D-dimer
Ferritin	Ferritin
Homocysteine	Homocysteine
IL18	Interleukin 18
IL1_RA	Interleukin1 receptor antagonist
Leptin	Leptin

Table 4.4: List of 60 blood pressure measures and biochemical assays (Continued)

Trait	Description
MPO	Myeloperoxidase
Adrenomedullin	Mid regional pro adrenomedullin
ANP	Mid regional pro atrial natriuretic peptide
Neopterin	Neopterin
BNP	N terminal prohormone of brain natriuretic peptide
PLA_A	Phospholipase A2 activity
PLA_M	Phospholipase A2 mass
PLGF	Placental growth factor
PON1	Paraoxonase 1
TIMP1	Tissue inhibitor metalloproteinase 1
Glucose	Glucose corrected for fasting
Insulin	Insulin corrected for fasting
Testerone	Testerone
Vitamin-B12	Vitamin B12 (Cobalamin)
Vitamin-D	Vitamin D

Table 4.5: All associations with discovery P<2e-04

			Discovery				Replication				Combined			
Trait	**Gene**	**Variant**	**N**	**Beta**	**SE**	**P-value**	**N**	**Beta**	**SE**	**P-value**	**N**	**Beta**	**SE**	**P-value**
Lp(a)	LPA	splice	6696	-0.608	0.031	2.17E-81	2200	-0.729	0.055	6.80E-39	8896	-0.637	0.027	1.53E-117
Vitamin-B12	FUT2	stop	6087	0.199	0.019	3.68E-26								
Galectin-3	TBPL2	stop	6648	-0.460	0.080	9.37E-09								
GCSF	ATP2C2	splice	6660	0.272	0.055	6.98E-07	2188	-0.037	0.105	7.25E-01	8848	0.206	0.049	2.27E-05
IL4	ATP2C2	splice	6660	0.258	0.055	2.48E-06	2188	0.035	0.105	7.42E-01	8846	0.209	0.061	5.91E-04
IFN-gamma	ATP2C2	splice	6660	0.255	0.055	3.24E-06	2188	0.060	0.105	5.72E-01	8841	0.051	0.016	1.45E-03
IL6	ATP2C2	splice	6660	0.251	0.055	4.58E-06	2188	0.073	0.105	4.88E-01	8848	0.213	0.049	1.16E-05
Endothelin1	FUT2	stop	6146	0.086	0.019	5.63E-06								
D-dimer	FGL1	frameshift	6582	0.210	0.046	6.12E-06								
IL12	ATP2C2	splice	6660	0.245	0.055	8.13E-06	2188	0.042	0.105	6.87E-01	8848	0.201	0.049	3.45E-05
IL17	ATP2C2	splice	6660	0.241	0.055	1.12E-05	2188	-0.136	0.105	1.95E-01	8848	0.160	0.049	9.91E-04
Systolic bp	ATP2C2	splice	25764	0.125	0.029	1.25E-05	9355	0.113	0.054	3.65E-02	35119	0.122	0.025	1.31E-06
IFN-gamma	P4HA3	stop	6655	0.080	0.019	1.70E-05	2186	-0.036	0.032	2.70E-01	8841	0.051	0.016	1.45E-03
IL17	P4HA3	stop	6655	0.080	0.019	1.72E-05	2186	0.015	0.033	6.34E-01	8841	0.064	0.016	7.27E-05
Vitamin-B12	CLYBL	stop	6600	-0.203	0.047	1.83E-05								

Table 4.5: All associations with discovery P<2e-04 (Continued)

			Discovery				Replication				Combined			
Trait	**Gene**	**Variant**	**N**	**Beta**	**SE**	**P-value**	**N**	**Beta**	**SE**	**P-value**	**N**	**Beta**	**SE**	**P-value**
Lp(a)	LPA	splice	6696	-0.608	0.031	2.17E-81	2200	-0.729	0.055	6.80E-39	8896	-0.637	0.027	1.53E-117
Vitamin-B12	FUT2	stop	6087	0.199	0.019	3.68E-26								
Galectin-3	TBPL2	stop	6648	-0.460	0.080	9.37E-09								
GCSF	ATP2C2	splice	6660	0.272	0.055	6.98E-07	2188	-0.037	0.105	7.25E-01	8848	0.206	0.049	2.27E-05
IL4	ATP2C2	splice	6660	0.258	0.055	2.48E-06	2188	0.035	0.105	7.42E-01	8846	0.209	0.061	5.91E-04
IFN-gamma	ATP2C2	splice	6660	0.255	0.055	3.24E-06	2188	0.060	0.105	5.72E-01	8841	0.051	0.016	1.45E-03
IL6	ATP2C2	splice	6660	0.251	0.055	4.58E-06	2188	0.073	0.105	4.88E-01	8848	0.213	0.049	1.16E-05
Endothelin1	FUT2	stop	6146	0.086	0.019	5.63E-06								
D-dimer	FGL1	frameshift	6582	0.210	0.046	6.12E-06								
IL12	ATP2C2	splice	6660	0.245	0.055	8.13E-06	2188	0.042	0.105	6.87E-01	8848	0.201	0.049	3.45E-05
IL17	ATP2C2	splice	6660	0.241	0.055	1.12E-05	2188	-0.136	0.105	1.95E-01	8848	0.160	0.049	9.91E-04
Systolic bp	ATP2C2	splice	25764	0.125	0.029	1.25E-05	9355	0.113	0.054	3.65E-02	35119	0.122	0.025	1.31E-06
IFN-gamma	P4HA3	stop	6655	0.080	0.019	1.70E-05	2186	-0.036	0.032	2.70E-01	8841	0.051	0.016	1.45E-03
IL17	P4HA3	stop	6655	0.080	0.019	1.72E-05	2186	0.015	0.033	6.34E-01	8841	0.064	0.016	7.27E-05
Vitamin-B12	CLYBL	stop	6600	-0.203	0.047	1.83E-05								
TNF-beta	HTRA4	frameshift	6669	-0.292	0.069	2.68E-05	2188	0.378	0.141	7.29E-03	8857	-0.172	0.062	5.54E-03

Table 4.5: All associations with discovery P<2e-04 (Continued)

			Discovery				Replication				Combined			
Trait	**Gene**	**Variant**	**N**	**Beta**	**SE**	**P-value**	**N**	**Beta**	**SE**	**P-value**	**N**	**Beta**	**SE**	**P-value**
IL4	ATP10B	stop	6673	0.186	0.045	3.55E-05	2189	0.141	0.086	1.01E-01	8862	0.177	0.040	9.71E-06
FGF	P4HA3	stop	6655	0.076	0.019	4.58E-05	2186	0.006	0.032	8.49E-01	8841	0.059	0.016	2.81E-04
TNF-beta	ATP10B	stop	6673	0.184	0.045	4.65E-05	2189	0.091	0.081	2.63E-01	8862	0.164	0.039	3.26E-05
TNF-beta	ATP2C2	splice	6660	0.223	0.055	4.69E-05	2188	-0.024	0.099	8.08E-01	8848	0.170	0.048	3.86E-04
IFN-gamma	ATP10B	stop	6673	0.183	0.045	4.81E-05	2189	0.037	0.086	6.68E-01	8862	0.152	0.040	1.45E-04
SDF1	ATP2C2	splice	6660	0.221	0.055	5.69E-05	2188	0.057	0.105	5.85E-01	8848	0.186	0.049	1.31E-04
TNF-beta	P4HA3	stop	6655	0.075	0.019	5.94E-05	2186	0.007	0.031	8.10E-01	8841	0.058	0.016	2.66E-04
FGF	ATP2C2	splice	6660	0.220	0.055	5.98E-05	2188	0.033	0.105	7.51E-01	8848	0.180	0.049	2.11E-04
IL18	EPPK1	frameshift	6677	-0.232	0.058	6.20E-05	2160	-0.240	0.102	1.88E-02	8837	-0.234	0.050	3.42E-06
PDGF	ATP10B	stop	6673	0.181	0.045	6.20E-05	2189	-0.011	0.086	9.02E-01	8862	0.139	0.040	4.84E-04
SDF1	ATP10B	stop	6673	0.178	0.045	7.62E-05	2189	-0.118	0.085	1.66E-01	8862	0.114	0.040	4.12E-03
Triglycerides	MS4A2	splice	25051	0.129	0.033	7.80E-05	9489	0.151	0.054	4.85E-03	34540	0.135	0.028	1.31E-06
VEGF	HTRA4	frameshift	6669	-0.274	0.069	7.86E-05	2188	-0.010	0.149	9.46E-01	8857	-0.227	0.063	3.10E-04
IL17	CLYBL	stop	6671	0.185	0.047	8.77E-05								
IL10	ATP2C2	splice	6660	0.214	0.055	9.33E-05	2188	0.134	0.105	2.04E-01	8848	0.197	0.049	5.08E-05
IL6	P4HA3	stop	6655	0.072	0.019	9.70E-05	2186	-0.016	0.033	6.17E-01	8841	0.051	0.016	1.73E-03

Table 4.5: All associations with discovery P<2e-04 (Continued)

			Discovery				Replication				Combined			
Trait	**Gene**	**Variant**	**N**	**Beta**	**SE**	**P-value**	**N**	**Beta**	**SE**	**P-value**	**N**	**Beta**	**SE**	**P-value**
IL17	HTRA4	frameshift	6669	-0.270	0.069	1.03E-04	2188	-0.008	0.149	9.57E-01	8857	-0.223	0.063	3.97E-04
IFN-gamma	CCL26	stop	6663	-0.307	0.080	1.15E-04	2192	-0.183	0.133	1.69E-01	8855	-0.274	0.068	5.94E-05
IL12	ATP10B	stop	6673	0.174	0.045	1.20E-04	2189	0.010	0.086	9.09E-01	8862	0.138	0.040	5.41E-04
IL6	ATP10B	stop	6673	0.173	0.045	1.23E-04	2189	0.024	0.086	7.79E-01	8862	0.141	0.040	4.15E-04
IL17	ATP10B	stop	6673	0.173	0.045	1.24E-04	2189	-0.003	0.086	9.74E-01	8862	0.135	0.040	7.13E-04
PDGF	P4HA3	stop	6655	0.071	0.019	1.31E-04	2186	-0.022	0.032	4.99E-01	8841	0.048	0.016	2.86E-03
GCSF	CLYBL	stop	6671	0.180	0.047	1.41E-04								
PDGF	ATP2C2	splice	6660	0.209	0.055	1.43E-04	2188	0.048	0.105	6.49E-01	8848	0.174	0.049	3.40E-04
GCSF	EFCAB3	stop	6606	0.157	0.042	1.86E-04	2192	-0.157	0.079	4.62E-02	8798	0.087	0.037	1.84E-02

Three of these association have been previously reported and represent positive controls for our approach: a strong association for the 2 splice variants (c.4974-2A>G and c.4289+1G>A) in the *Lipoprotein(a)* gene (*LPA*) with lipoprotein(a) measurements in plasma ($P_{discovery}$ = 2.17×10^{-81}, $P_{discovery+replication}$ = 1.53×10^{-117}, combined $\hat{\beta}$ = -0.64 or -8.77 mg/dL per allele), the W154X variant in *Fucosyltransferase 2* (*FUT2*) with increased Vitamin B12 levels [24] ($\hat{\beta}$ = 0.2, P = 3.7×10^{-26} or 43.38 pg/mL per allele) and the R225X variant in the *Citrate Lyase Beta Like* gene (*CLYBL*) with decreased Vitamin B12 levels [25] ($\hat{\beta}$ = -0.2, P = 1.8×10^{-5} or -43.38 pg/mL per allele) [26]. The boxplots for these associations are shown in Figure 4.6.

In addition to an extremely strong correlation between lipoprotein(a) levels and cardiovascular disease, it has been previously reported that genetic variants that elevate circulating lipoprotein(a) levels are cardiovascular risk factors [27,28]. The converse, critical for evaluation of the therapeutic hypothesis of inhibition, that lowering lipoprotein(a) levels can confer cardiovascular protection has not yet been evaluated. With access to centralized medical records available to research in Finland, we utilized the strong lipoprotein(a) lowering variants discovered here to evaluate the impact of lipoprotein(a) lowering via Mendelian randomization. Using a Cox proportional hazards model for incident cardiovascular disease in these cohorts (adjusted for age, gender and therapies), the composite *LPA* variant was found to protect against coronary heart disease (Hazard Ratio HR = 0.79, P = 6.7×10^{-3}), demonstrating that lowering lipoprotein(a) levels are likely to confer protection for cardiovascular diseases. We confirmed this finding using three independent non-Finnish datasets: an early onset myocardial infarction dataset of 18,000 individuals and two studies from the Estonian Biobank (4,600 and 7,953 individuals respectively), which collectively replicated the observation that the *LPA* variants confer cardioprotective effect (OR = 0.87, P = 0.016).

Figure 4.6: Boxplots for the known and novel associations

The normalized Z-scores are shown for the homozygous alternates, heterozygous and homozygous reference individuals on the left, while the log of the unnormalized values are shown on the right for the 3 genotypes.

Figure 4.6: Boxplots for the known and novel associations (Continued)

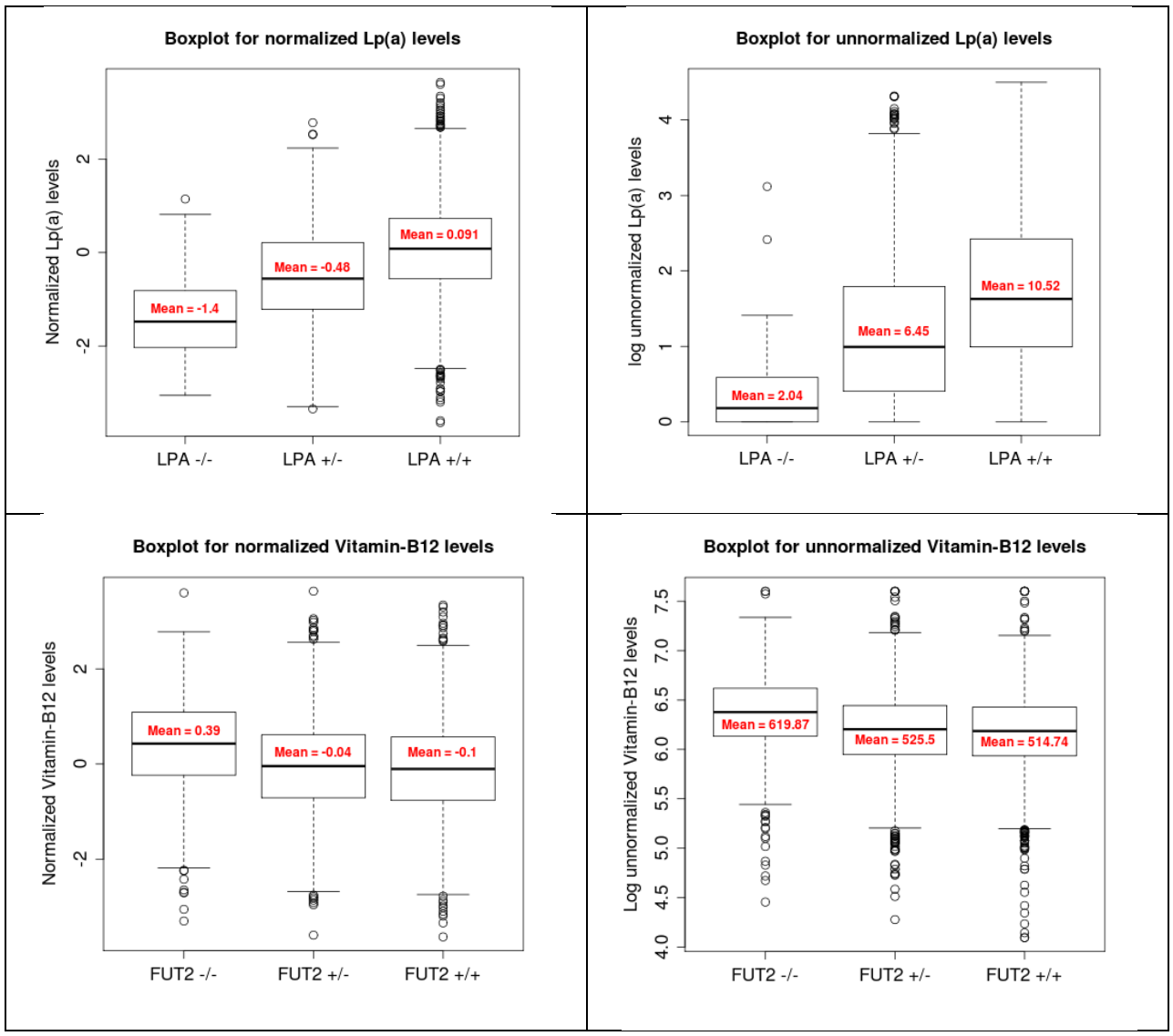


Figure 4.6: Boxplots for the known and novel associations (Continued)

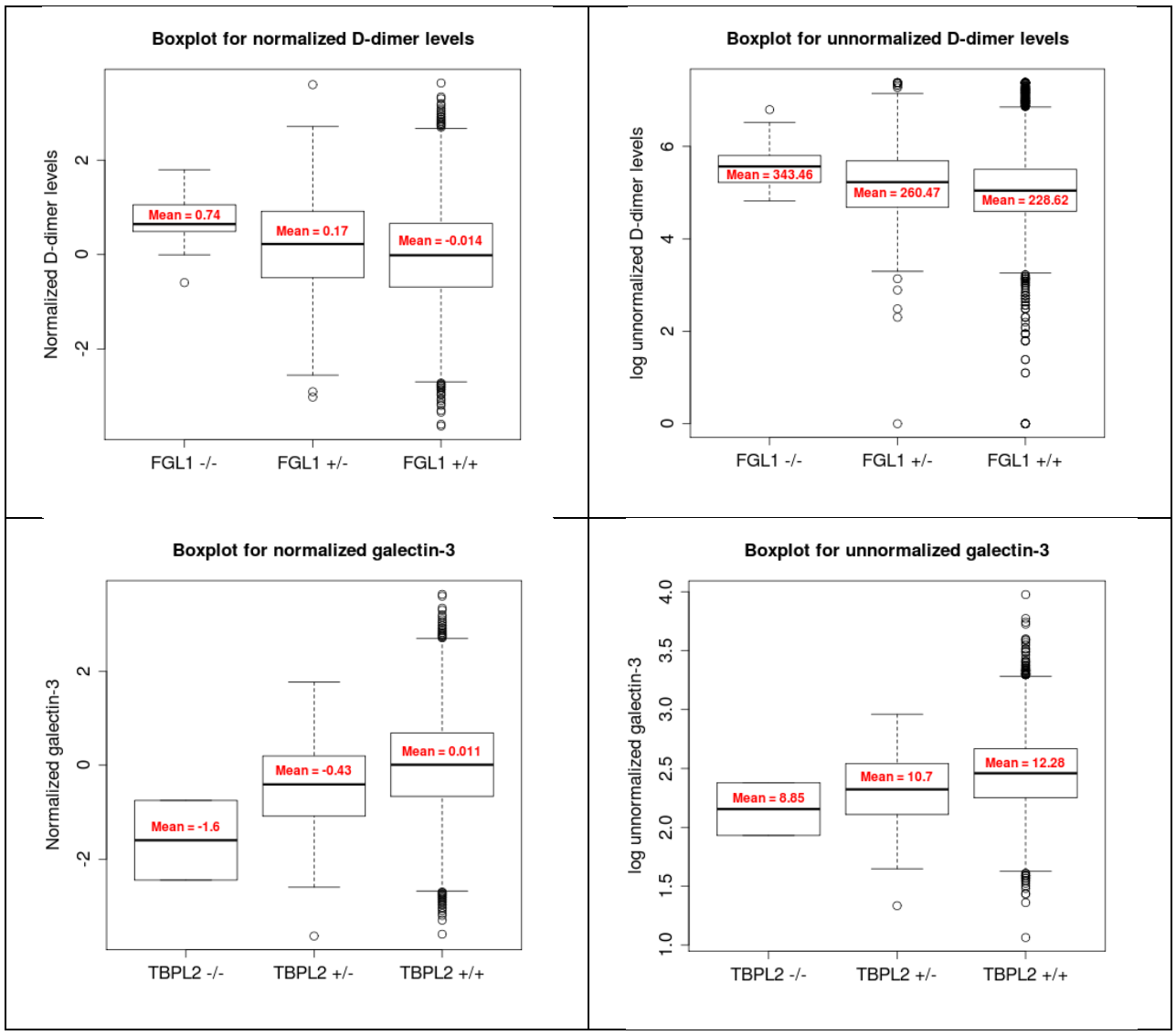
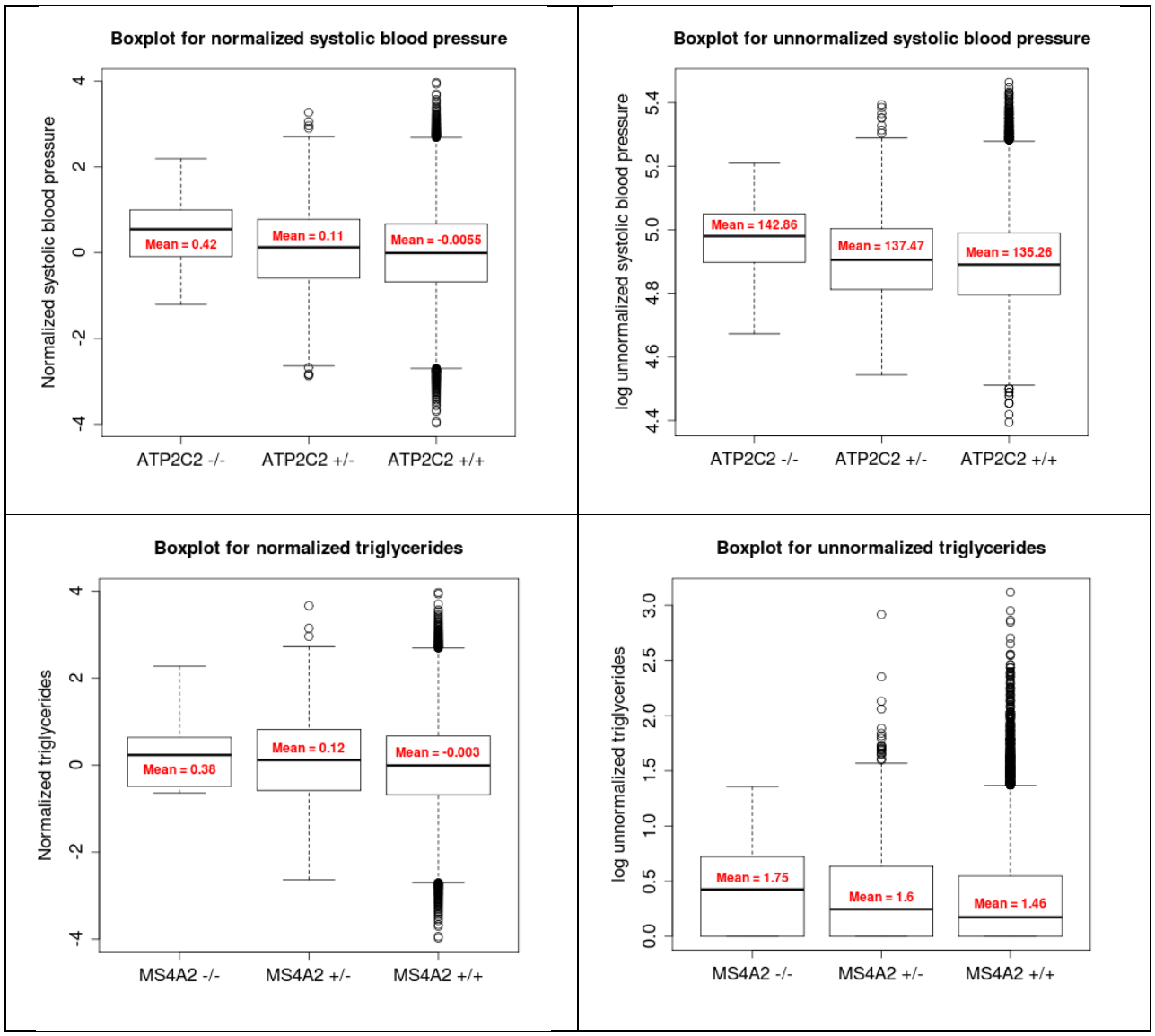


Figure 4.6: Boxplots for the known and novel associations (Continued)



After meta-analyzing all the datasets, the final odds ratio was found to be 0.84 ($P = 3\times10^{-4}$, Figure 4.7). Similar to observations previously reported for *PCSK9* (which induces cardioprotection through the lowering of low density lipoprotein levels) [29], we found 227 individuals who are homozygous or compound heterozygous for the two *LPA* splice variants with no evidence for increased morbidity or mortality based on National Health Records. This suggests that reduction of lipoprotein(a) is well-tolerated and might constitute a potential drug target for cardiovascular diseases (Table 4.6).

In addition, we observed novel associations were observed for the *TBPL2*, *FGL1, MS4A2* and *ATP2C2* variants. The R331X variant in the *TATA Box Binding Protein Like 2* gene (*TBPL2*) was associated to decreased levels of galectin-3 ($\hat{\beta}$ = -0.46, $P = 9.4\times10^{-9}$ or -1.77 ng/dL per allele). Increased galectin-3 levels have previously been associated with increased risk for heart failure, chronic kidney disease and diabetes [30,31,32,33]. The *TBPL2* gene encodes a widely expressed transcription factor that has been described to be involved in myoblast differentiation [34] and *Tbpl2*$^{-/-}$ mice and zebrafish have defects in reproduction and haematopoiesis [35,36]. We did not observe any association between *TBPL2* R311X and cardiovascular or diabetes outcomes.

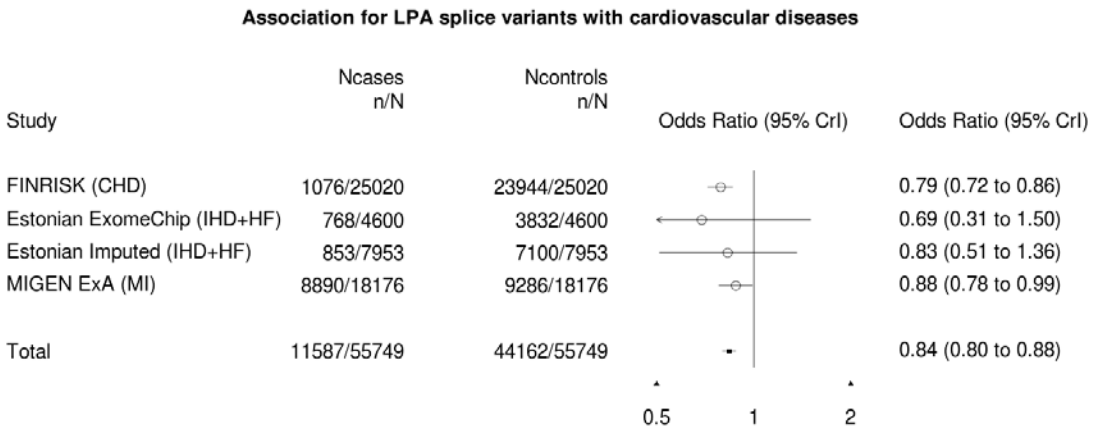


Figure 4.7: Forest plot for the *LPA* splice variants with cardiovascular diseases.

The cardiovascular diseases are defined as coronary heart disease (CHD), ischemic heart disease (IHD), heart failure (HF) or myocardial infarction (MI).

Table 4.6: Correlations between the combined *LPA* variant and various disease states

The rows with significant correlation between the levels of the biomarker and disease status (P<1e-03) are shaded in blue and the rows with significant association (P<=0.05) between the variant and disease status (allelic or homozygous tests) are highlighted in red text.

	T-test		Association of combined *LPA* variant with Lp(a) levels adjusted for disease			Association of combined *LPA* variant with disease				Association of combined *LPA* homozygotes with disease				
Disease	**Higher Lp(a) levels**	**Lower Lp(a) levels**	**N**	**Beta**	**P-value**	**N cases**	**N controls**	**Beta**	**P-value**	**Case aa**	**Case Aa**	**Case AA**	**OR**	**P-value**
Cardio vascular disease	6.73E-04	9.99E-01	6696	-0.607	3.03E-81	2272	23508	-0.112	7.73E-02	12	302	1958	0.77	4.99E-01
Coronary heart disease	8.19E-04	9.99E-01	6696	-0.607	2.81E-81	1645	24135	-0.153	3.86E-02	6	215	1424	0.53	1.56E-01
Myocardial infarction	1.04E-02	9.90E-01	6696	-0.608	1.91E-81	953	24827	-0.148	1.18E-01	3	125	825	0.46	2.23E-01
Acute coronary syndrome	1.23E-02	9.88E-01	6696	-0.608	2.20E-81	1286	24494	-0.111	1.70E-01	6	171	1109	0.69	4.81E-01
Stroke	3.72E-02	9.63E-01	6696	-0.608	2.21E-81	1038	24742	-0.048	5.80E-01	8	140	890	1.16	6.94E-01
Major adverse cardiac events	4.30E-02	9.57E-01	6696	-0.608	2.12E-81	3207	22573	-0.057	2.87E-01	20	442	2745	0.93	8.17E-01
Ischemic heart disease	3.02E-01	6.98E-01	6696	-0.609	1.71E-81	2757	23023	-0.009	8.78E-01	19	391	2347	1.04	9.01E-01
Cancer	6.43E-01	3.57E-01	6696	-0.608	2.33E-81	1934	23846	0.079	2.00E-01	14	296	1624	1.09	7.70E-01
Rheumatoid arthritis	9.13E-01	8.68E-02	6696	-0.608	2.83E-81	1149	24631	0.015	8.53E-01	10	162	977	1.33	3.53E-01
Asthma	9.40E-01	5.97E-02	6696	-0.608	2.40E-81	3137	22643	0.025	6.11E-01	19	460	2658	0.90	7.26E-01
Heart failure	9.81E-01	1.92E-02	6696	-0.608	2.86E-81	1557	24223	0.044	5.26E-01	14	226	1317	1.38	2.58E-01
High blood pressure	9.98E-01	2.00E-03	6696	-0.607	4.07E-81	5197	20583	0.070	9.71E-02	47	757	4393	1.49	2.20E-02
Diabetes	1.00E+00	5.27E-07	6696	-0.607	2.12E-81	2866	22914	0.030	5.70E-01	22	414	2430	1.17	4.65E-01

The 1-bp c.545_546insA frameshift in the *Fibrinogen-like 1* gene (*FGL1*) was associated with increased D-dimer levels ($\hat{\beta}$ = 0.21, P = 6.1×10^{-6} or 52.23 ng/mL per allele). D-dimers are products of fibrin degradation and their concentration in the blood flow is clinically used to monitor thrombotic activity. The role of *FGL1* in clot formation remains unclear: although *FGL1* is homologous with fibrinogen, it lacks the essential structures for fibrin formation, with one study suggesting its presence in fibrin clots [37]. In addition, given prior links between variants associated with D-dimer levels and stroke, we utilized the same Mendelian randomization approach as for *LPA* above and found a nominally significant association between *FGL1* c.545_546insA and increased risk of ischemic stroke (OR = 1.32, P = 0.024). If replicated, this would be consistent with modest risk increase for stroke that other variants associated to circulating D-dimer levels, such as reported for variants in coagulation *Factor V*, *Factor III* and *FGA* [38].

We found suggestive associations for the c.637-1G>A splice variant in the *membrane-spanning 4-domains, subfamily A, member 2* gene (*MS4A2*) with triglycerides ($P_{discovery}$ = 7.80×10^{-5}, $P_{discovery+replication}$ = 1.31×10^{-6}, $\hat{\beta}$ = 0.14 or 0.14 mmol/L per allele). This observation is consistent with our previously published study of 631 individuals in the DILGOM subset of FINRISK showing that whole blood expression of *MS4A2* was strongly negatively associated with total triglycerides ($\hat{\beta}$ = -1.62, P = 2.1×10^{-27}) [39] and a wide range of systemic metabolic traits[40]. A similar but insignificant trend was observed in 15,696 individuals from the D2D2007, DPS, FUSION, METSIM and DRSEXTRA cohorts ($\hat{\beta}$ = 0.04, P = 0.32). The *MS4A2* gene encodes the β-subunit of the high affinity IgE receptor, a key mediator of the acute phase inflammatory response.

The c.2482-2A>C splice variant in the *ATPase Ca++ Transporting Type 2C Member 2* gene (*ATP2C2*) was associated with increased systolic blood pressure ($P_{discovery} = 1.25\times10^{-5}$, $P_{discovery+replication} = 1.3\times10^{-6}$, $\hat{\beta} = 0.12$ or 2.13 mmHg per allele (an association that is undisturbed by correction for lipid lowering medication ($\hat{\beta} = 0.12$, $P = 1.75\times10^{-5}$) or blood pressure lowering medication ($\hat{\beta} = 0.13$, $P = 1.3\times10^{-5}$)). Based on its structure, *ATP2C2* is predicted to catalyze the hydrolysis of ATP coupled with calcium transport. Interestingly, the *ATP2C2* c.2482-2A>C variant is also significantly associated to several highly correlated immune markers, such as granulocyte colony-stimulating factor ($\hat{\beta} = 0.26$, $P = 6.98\times10^{-7}$), interleukin-4 ($\hat{\beta} = 0.27$, $P = 2.48\times10^{-6}$), interferon-γ ($\hat{\beta} = 0.26$, $P = 3.24\times10^{-6}$) and interleukin-6 ($\hat{\beta} = 0.25$, $P = 4.58\times10^{-6}$).

DISCUSSION

In this study, both replicated results and novel associations demonstrate the association of low-frequency LoF variants with various complex traits and diseases. In addition, we discovered a novel cardiovascular protective effect from splice variants in the *LPA* gene, suggesting that knocking down levels of circulating Lp(a) can confer a protection from cardiovascular diseases. Given that we detected numerous individuals in these adult population cohorts, healthy and in the expected Hardy-Weinberg proportions, carrying a complete knockout of *LPA* (homozygous or compound heterozygous for the 2 splice variants), this suggests that knocking out the gene in humans does not result in severe medical consequences. As such, this study provides a substantial body of human data that *LPA* may be an effective target for therapeutic purposes.

As more Finnish samples are being sequenced, these enriched variants can also be imputed with high precision to the large number of existing samples with array-based GWAS genotypes. This advantage is likely to be more pronounced for the much larger pool of missense

variation – while one can presume all LoF variants in a gene might have a comparable effect on phenotype (and thereby burden tests of LoF variants in an out-bred sample is not at a great disadvantage compared to isolated populations), it is evident that many rare missense variants within the same gene will not all have the same impact on gene function. Thus the ability to assess single low-frequency variants conclusively, especially since they will include an excess of damaging variants enriched through a bottleneck, rather than perform burden tests on heterogeneous sets of extremely rare variants, will offer substantial ongoing advantage to isolated population studies as in recent findings. The Finnish population, with the concomitant advantage of existing genome-wide profiling of tens of thousands of individuals and numerous bio-repositories integrated with population-wide medical registry information, is well-positioned to be at the forefront of this new wave of discovery.

MATERIALS AND METHODS

Exome sequencing quality control, annotation and filtering

Raw Binary Sequence Alignment/Map (BAM) files from the various projects were jointly processed at the Broad Institute and joint variant calling was performed on all exomes to minimize batch differences. We annotated variants using a custom pipeline (MacArthur *et al.*, unpublished) and we required all variants to pass the basic GATK filters and required all genotypes to have a quality score of ≥30, read depth of ≥10 and allele balance of between 0.3 and 0.7 for heterozygous calls and <0.1 for homozygous calls. Allele counts and frequencies were calculated within the 3,000 individuals for Finns and NFEs respectively.

Sequenom genotyping

Genotyping was performed using the iPLEX™ Gold Assay (Sequenom® Inc.). Assays for all SNPs were designed using the eXTEND suite and MassARRAY Assay Design software version 3.1 (Sequenom® Inc.). Amplification was performed in a total volume of 5μL containing ~10ng genomic DNA, 100nM of each PCR primer, 500μM of each dNTP, 1.25 x PCR buffer (Qiagen), 1.625mM $MgCl_2$and 1U HotStar Taq® (Qiagen). Reactions were heated to 94 °C for 15 min followed by 45 cycles at 94 °C for 20 s, 56 °C for 30 s and 72 °C for 1 min, then a final extension at 72 °C for 3 min. Unincorporated dNTPs were SAP digested prior to iPLEX™ Gold allele specific extension with mass-modified ddNTPs using an iPLEX Gold reagent kit (Sequenom® Inc.). SAP digestion and extension were performed according to the manufacturer's instructions with reaction extension primer concentrations adjusted to between 0.7-1.8μM, dependent upon primer mass. Extension products were desalted and dispensed onto a SpectroCHIP using a MassARRAY Nanodispenser prior to MALDI-TOF analysis with a

MassARRAY Analyzer Compact mass spectrometer. Genotypes were automatically assigned and manually confirmed using MassARRAY TyperAnalyzer software version 4.0 (Sequenom® Inc.). The variants were then checked for concordance in allele frequencies with the exome sequencing data.

Phenotyping

Data on disease status from National Health registers (Hospital Discharged Registers maintained by THL (Institute for Health and Welfare, Finland), Cause of Death Register, Statistics Finland and Prescription Medication Register, THL) for FINRISK, Health2000 and the Young Finns Study participants of this study were collected and curated. A description of each cohort is provided in the Supplement.

Analyses of RNA sequencing data

To analyze the effects of the LoF variants on gene expression, we used RNA sequencing data from two major studies: the GEUVADIS project [19] with RNA sequencing data from lymphoblastoid cell lines of 462 individuals participants from the 1000 Genomes Project [41]), and the GTEx project with RNA-sequencing data from a total of 175 individuals with 1-30 tissues each (http://www.broadinstitute.org/gtex/) [20]. The processing of the GEUVADIS data and the methods for allele-specific expression analysis are described in Lappalainen *et al.* [19] and the GTEx data were analyzed using similar methods. Allele-specific expression analysis was used primarily to capture nonsense-mediated decay. Additionally, to assess whether LoF variants lead to decreased exon expression levels overall or for individual exons, we calculated an empirical p-value for each exon of all the LoF genes with respect to all other exons genome-

wide, denoting the proportion of all exons where carriers of the LoF variants are more extreme than in the each studied exon in LoF variant genes. The analyses were performed separately in each studied tissue: lymphoblastoid cell lines from the GEUVADIS data and nine tissues from the GTEx data. The significance threshold after correcting for the total number of tested exons across all tissues is $0.05/1070 = 4.67\times10^{-5}$.

Statistical analyses and methods

Inverse rank-based normalization was performed on the quantitative measurements in males and females separately, with linear regression residuals using age and age^2 as covariates. Linear regression was then performed on the normalized Z-scores using R to obtain the statistics for the associations. We tested the correlations between the quantitative measurements and disease outcomes using two one-tailed t-tests to assess the significance of observing higher levels of the quantitative measurements in cases (individuals with the disease outcomes) versus controls (individuals without the disease outcomes), as well as lower levels of the quantitative measurements in cases versus controls. To test the association of the variants with the prevalent disease outcomes, we performed a logistic regression in R to obtain the reported statistics. In addition, a Fisher's Exact Test on the homozygous counts in cases and controls were performed to test for association with the homozygotes. The results for the LPA with cardiovascular disease association from MIGen ExA and the Estonian Biobank were meta-analyzed using METAL [42] and the combined results with FINRISK were obtained using the Fisher's Combined P method with 4 degrees of freedom.

Associations between *MS4A2* c.637-1G>A, gene expression and triglycerides

We fit a linear model in which the $\log_2$-normalised gene probe expression of individual *i* was regressed on the LoF genotype, which was encoded as $X_i = 0$, 1 or 2 for the LoF genotypes -/-, +/- or +/+ respectively and association analysis of *MS4A2* gene expression and triglycerides was performed as previously reported [39]. Briefly, we used a multivariate linear regression adjusted for age, gender, and use of cholesterol or blood pressure lowering medication. We further tested for association between *MS4A2* c.637-1G>A and triglycerides using a 2-sided t-test.

ACKNOWLEDGEMENTS

We acknowledge support and funding from the following grants: the European Commission FP7 projects no. 201413 ENGAGE (to A.P.), project no. 242167 SynSys (to A.P.), Health-2010 -projects no. 261433 BioSHare (to A.P.) and project no. 261123 gEUVADIS (to A.P.), the Academy of Finland grants no. 251704 and 263401 (to A.P.), the Finnish Foundation for Cardiovascular Research (to A.P.), the Sigrid Juselius Foundation (to A.P.), NIH/RFA-HL-12-007 (to A.P.), the European Commission Health-2010-project no. 261433 BioSHare (to S.R.), the Academy of Finland grants no 255847 and no 251217 (to S.R.), the Finnish Foundation for Cardiovascular Research (to S.R.), the Sigrid Juselius Foundation (to S.R.), the Academy of Finland grant #139635 (to V.S.), the Finnish Foundation for Cardiovascular Research (to V.S.), the Australian National Health and Medical Research Council Early Career Fellowship no. 637400 (to M.I.), the Wellcome Trust Research Career Development Fellow 086596/Z/08/Z (to C.M.L.), GoT2D RC2-DK088389 (to D.M.A), GoT2D Wellcome Trust 090367 (to M.McC.), Wellcome Trust 098381 (to M.McC.), T2DGENES NIDDM U01-DK-085545 (to M.McC.), DK062370 (to M.B.), DK085584 (to M.B.), (to M.B.), DK088389 (to M.B.), Academy of Finland grants 141054, 265240, 263278 (to J.K.), Academy of Finland grant 250422 (to P.W.), the European Union's Seventh Framework Programme FP7/2007-2013 [HEALTH-F2-2011-278913, BiomarCaRE], the Targeted Financing from the Estonian Ministry of Science and Education SF0180142s08 (to A.M.), the US National Institute of Health R01DK075787 (to J.N.H., A.M.), the Development Fund of the University of Tartu grant SP1GVARENG (to A.M.), the European Regional Development Fund to the Centre of Excellence in Genomics (EXCEGEN) grant 3.2.0304.11-0312 (to A.M.), and FP7 grant 313010 (to A.M.). We thank the GoT2D, T2D-GENES and MIGen ExA groups for providing replication data, as well as the

GTEx Consortium and Geuvadis Consortium for the use of the RNA sequencing data, and NHLBI GO Exome Sequencing Project and its ongoing studies which produced and provided exome variant calls for comparison: the Lung GO Sequencing Project (HL-102923), the WHI Sequencing Project (HL-102924), the Broad GO Sequencing Project (HL-102925), the Seattle GO Sequencing Project (HL-102926) and the Heart GO Sequencing Project (HL-103010). Exome Chip genotyping in the Myocardial infarction Genetics Exome Array Consortium (MIGen ExA) was supported by NIH RC2 HL-102925 (to S.G. and D.M.A.) and an investigator-initiated research grant from Merck to S.Kathiresan.

COMPETING FINANCIAL INTERESTS

Abbott Diagnostics provided test reagents for FINRISK 1997 determinations of Galectin-3, Lp(a) and D-dimer within the framework of the MORGAM Biomarker Study and the BiomarCaRE project. S.Blankenberg has received honoraria from Abbott Diagnostics, SIEMENS, Thermo Fisher and Roche Diagnostics and is a consultant for Thermo Fisher. V.S. has received a speaker honorarium from Roche Diagnostics. All other co-authors reported no conflicts of interest.

REFERENCES

1. Pollin TI, Damcott CM, Shen H, Ott SH, Shelton J, et al. (2008) A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. Science 322: 1702-1705.

2. Huyghe JR, Jackson AU, Fogarty MP, Buchkovich ML, Stancakova A, et al. (2012) Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. Nat Genet.

3. Styrkarsdottir U, Thorleifsson G, Sulem P, Gudbjartsson DF, Sigurdsson A, et al. (2013) Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits. Nature 497: 517-520.

4. Jonsson T, Atwal JK, Steinberg S, Snaedal J, Jonsson PV, et al. (2012) A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. Nature 488: 96-99.

5. Bevilacqua L, Doly S, Kaprio J, Yuan Q, Tikkanen R, et al. (2010) A population-specific HTR2B stop codon predisposes to severe impulsivity. Nature 468: 1061-1066.

6. Gudmundsson J, Sulem P, Gudbjartsson DF, Masson G, Agnarsson BA, et al. (2012) A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. Nat Genet 44: 1326-1329.

7. Aminoff M, Carter JE, Chadwick RB, Johnson C, Grasbeck R, et al. (1999) Mutations in CUBN, encoding the intrinsic factor-vitamin B12 receptor, cubilin, cause hereditary megaloblastic anaemia 1. Nat Genet 21: 309-313.

8. Aaltonen J, Bjorses P (1999) Cloning of the APECED gene provides new insight into human autoimmunity. Ann Med 31: 111-116.

9. Savukoski M, Klockars T, Holmberg V, Santavuori P, Lander ES, et al. (1998) CLN5, a novel gene encoding a putative transmembrane protein mutated in Finnish variant late infantile neuronal ceroid lipofuscinosis. Nat Genet 19: 286-288.

10. de la Chapelle A, Wright FA (1998) Linkage disequilibrium mapping in isolated populations: the example of Finland revisited. Proc Natl Acad Sci U S A 95: 12416-12423.

11. Polvi A, Linturi H, Varilo T, Anttonen AK, Byrne M, et al. (2013) The Finnish Disease Heritage Database (FinDis) update - a database for the genes mutated in the Finnish Disease Heritage brought to the next-generation sequencing era. Hum Mutat.

12. Vartiainen E, Laatikainen T, Peltonen M, Juolevi A, Mannisto S, et al. (2010) Thirty-five-year trends in cardiovascular risk factors in Finland. Int J Epidemiol 39: 504-518.

13. Raitakari OT, Juonala M, Ronnemaa T, Keltikangas-Jarvinen L, Rasanen L, et al. (2008) Cohort profile: the cardiovascular risk in Young Finns Study. Int J Epidemiol 37: 1220-1226.

14. Smeitink JA, Elpeleg O, Antonicka H, Diepstra H, Saada A, et al. (2006) Distinct clinical phenotypes associated with a mutation in the mitochondrial translation elongation factor EFTs. Am J Hum Genet 79: 869-877.

15. Vedrenne V, Galmiche L, Chretien D, de Lonlay P, Munnich A, et al. (2012) Mutation in the mitochondrial translation elongation factor EFTs results in severe infantile liver failure. J Hepatol 56: 294-297.

16. Amberger J, Bocchini CA, Scott AF, Hamosh A (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res 37: D793-796.

17. Meetei AR, Medhurst AL, Ling C, Xue Y, Singh TR, et al. (2005) A human ortholog of archaeal DNA repair protein Hef is defective in Fanconi anemia complementation group M. Nat Genet 37: 958-963.

18. Singh TR, Bakker ST, Agarwal S, Jansen M, Grassman E, et al. (2009) Impaired FANCD2 monoubiquitination and hypersensitivity to camptothecin uniquely characterize Fanconi anemia complementation group M. Blood 114: 174-180.

19. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, Monlong J, et al. (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature.

20. Consortium G (2013) The Genotype-Tissue Expression (GTEx) project. Nat Genet 45: 580-585.

21. Harris BE, Carpenter JT, Diasio RB (1991) Severe 5-fluorouracil toxicity secondary to dihydropyrimidine dehydrogenase deficiency. A potentially more common pharmacogenetic syndrome. Cancer 68: 499-501.

22. Enns GM, Barkovich AJ, van Kuilenburg AB, Manning M, Sanger T, et al. (2004) Head imaging abnormalities in dihydropyrimidine dehydrogenase deficiency. J Inherit Metab Dis 27: 513-522.

23. Van Kuilenburg AB, Vreken P, Abeling NG, Bakker HD, Meinsma R, et al. (1999) Genotype and phenotype in patients with dihydropyrimidine dehydrogenase deficiency. Hum Genet 104: 1-9.

24. Hazra A, Kraft P, Selhub J, Giovannucci EL, Thomas G, et al. (2008) Common variants of FUT2 are associated with plasma vitamin B12 levels. Nat Genet 40: 1160-1162.

25. Lin X, Lu D, Gao Y, Tao S, Yang X, et al. (2012) Genome-wide association study identifies novel loci associated with serum level of vitamin B12 in Chinese men. Hum Mol Genet 21: 2610-2617.

26. Grarup N, Sulem P, Sandholt CH, Thorleifsson G, Ahluwalia TS, et al. (2013) Genetic architecture of vitamin B12 and folate levels uncovered applying deeply sequenced large datasets. PLoS Genet 9: e1003530.

27. Clarke R, Peden JF, Hopewell JC, Kyriakou T, Goel A, et al. (2009) Genetic variants associated with Lp(a) lipoprotein level and coronary disease. N Engl J Med 361: 2518-2528.

28. Kamstrup PR, Tybjaerg-Hansen A, Steffensen R, Nordestgaard BG (2009) Genetically elevated lipoprotein(a) and increased risk of myocardial infarction. JAMA 301: 2331-2339.

29. Cohen J, Pertsemlidis A, Kotowski IK, Graham R, Garcia CK, et al. (2005) Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nat Genet 37: 161-165.

30. Ho JE, Liu C, Lyass A, Courchesne P, Pencina MJ, et al. (2012) Galectin-3, a marker of cardiac fibrosis, predicts incident heart failure in the community. J Am Coll Cardiol 60: 1249-1256.

31. van der Velde AR, Gullestad L, Ueland T, Aukrust P, Guo Y, et al. (2013) Prognostic value of changes in galectin-3 levels over time in patients with heart failure: data from CORONA and COACH. Circ Heart Fail 6: 219-226.

32. O'Seaghdha CM, Hwang SJ, Ho JE, Vasan RS, Levy D, et al. (2013) Elevated Galectin-3 Precedes the Development of CKD. J Am Soc Nephrol.

33. Weigert J, Neumeier M, Wanninger J, Bauer S, Farkas S, et al. (2010) Serum galectin-3 is elevated in obesity and negatively correlates with glycosylated hemoglobin in type 2 diabetes. J Clin Endocrinol Metab 95: 1404-1411.

34. Deato MD, Tjian R (2007) Switching of the core transcription machinery during myogenesis. Genes Dev 21: 2137-2149.

35. Gazdag E, Santenard A, Ziegler-Birling C, Altobelli G, Poch O, et al. (2009) TBP2 is essential for germ cell development by regulating transcription and chromatin condensation in the oocyte. Genes Dev 23: 2210-2223.

36. Hart DO, Raha T, Lawson ND, Green MR (2007) Initiation of zebrafish haematopoiesis by the TATA-box-binding protein-related factor Trf3. Nature 450: 1082-1085.

37. Rijken DC, Dirkx SP, Luider TM, Leebeek FW (2006) Hepatocyte-derived fibrinogen-related protein-1 is associated with the fibrin matrix of a plasma clot. Biochem Biophys Res Commun 350: 191-194.

38. Smith NL, Huffman JE, Strachan DP, Huang J, Dehghan A, et al. (2011) Genetic predictors of fibrin D-dimer levels in healthy adults. Circulation 123: 1864-1872.

39. Inouye M, Silander K, Hamalainen E, Salomaa V, Harald K, et al. (2010) An immune response network associated with blood lipid levels. PLoS Genet 6: e1001113.

40. Inouye M, Kettunen J, Soininen P, Silander K, Ripatti S, et al. (2010) Metabonomic, transcriptomic, and genomic variation of a population cohort. Mol Syst Biol 6: 441.

41. MacArthur DG, Balasubramanian S, Frankish A, Huang N, Morris J, et al. (2012) A systematic survey of loss-of-function variants in human protein-coding genes. Science 335: 823-828.

42. Willer CJ, Li Y, Abecasis GR (2010) METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26: 2190-2191.

CHAPTER 5

Concluding Remarks

OVERVIEW

Whole-exome and whole-genome sequencing has revolutionized the way we map genes and variants in Mendelian and complex diseases in the past few years. In the near future, these technologies might be routinely used for both research as well as applied clinical diagnosis to identify, understand and treat genetic causes for various diseases. As approaches and methods for discovering causal or risk variants mature, we can expect the interpretation of these variants to remain an important challenge for human disease genetics.

MAJOR FINDINGS

Chapter 2

In this chapter, we discovered a small but significant 5% contribution to autism risk from rare recessively-acting variants on the autosomes and X-chromosome. The excess was mainly driven by genes that found to be expressed in the brain. While we found some genes that were previously implicated in Mendelian diseases, there was no evidence for specific genes.

Chapter 3

In this chapter, we discovered 3 families with autism and intellectual disability with rare homozygous missense mutations in the *DHCR24* gene. We developed a novel statistical method to evaluate the significance for such an observation in outbred and consanguineous populations. Finally, we adapted a yeast biochemical assay to evaluate the efficiency of desmosterol to cholesterol synthesis for these missense variants to understand the functionality and role of these rare missense variants in autism and intellectual disability.

Chapter 4

In this chapter, we explored the genetic architecture of the Finnish population versus other European populations and found that there are proportionally more deleterious rare and low-frequency variants in Finns. We genotyped 83 loss-of-function low-frequency variants in a large number of Finnish samples and associated these variants with 60 biochemical measurements and traits. In doing so, we discovered a strong association between splice variants in *LPA* with decreased circulating lipoprotein(a) levels and this translated to protection against cardiovascular heart disease.

FUTURE DIRECTIONS

With the decreasing costs of whole-genome and whole-exome sequencing, genetic mapping in human diseases has been revolutionized since 2009. The interpretation of genetic variation in the human genome remains a challenge, even for an individual genome [1]. While the interpretation of predicted loss-of-function variants in the human exome is easier, such variation comprise of a tiny proportion of the whole human exome. The discovery of missense and synonymous variation from healthy and diseased human exomes will no doubt be the norm rather than abnormality. As such, it would seem that rapid functional assays to allow the assessment and interpretation of such variants with unknown significance will be increasingly important. In addition, as such whole-genome and whole-exome sequencing become the norm for clinical diagnoses, it will be extremely important to develop and adapt new technologies that can allow for a rapid and unbiased way of assessing the “functionality” of such variants.

Rapid assessment of functionality of human coding variation

As we have learnt from our *DHCR24* project described in Chapter 3, such functionality assays are not necessarily straightforward or obvious, even for a cholesterol gene. For instance, although *DHCR24* is involved in the synthesis of cholesterol from desmosterol, this might not be the best or most suitable assay for *DHCR24* in some instances. The gene has also been demonstrated to play a role in regulating the oxidative stress response pathway by binding to p53 and another "functionality" assay that might be suitable for studying the role of variation in *DHCR24* can involve testing the strength of p53 binding rather than cholesterol synthesis.

Moreover, a genome-wide "functionality" assay to probe all variation across all genes will be extremely difficult to design. One genomics approach that has been suggested is the use of high-throughput RNA sequencing to correlate the missense variations with gene expression. This approach might work in evaluating which loss-of-function variants result in reduced gene expression as a consequence of nonsense-mediated decay. However, if the missense variants result in decreased protein interaction, DNA binding or activation of a second messenger within the cell, this will not necessarily translate to an interpretable readout of the functionality of such missense variants.

Non-coding variation in the human genome

Understanding the role of non-coding variation in the human genome from whole-genome sequencing will prove to be extremely difficult but potentially important. As discussed earlier, the yield of genetic discoveries in 250 clinical exomes is approximately 25% [2], suggesting that the remaining 75% have yet to be discovered, either from coding variants of unknown significance, non-coding variants or other unusual modes of inheritance. One of the

earliest papers that have explored this question is by Weedon *et al*. where the authors discovered 6 recessive mutations in ten families with pancreatic agenesis in a ~400 bp enhancer region downstream of *PTF1A* [3]. Such studies are extremely difficult as they involve interrogating the 99% non-coding regions of the human genome. However, we might expect human disease genetics to move rapidly into this area as we become more effective in understanding coding variation in the human genome.

Unusual modes of inheritance

As we become more efficient in our discovery of rare disease-causing or disease-association variants in Mendelian and complex diseases, we can expect an increasing amount of research into unusual modes of inheritance that can aid in our understanding of disease etiologies. A recent paper described the discovery of a recessive allele (Arg229Gln) in the *NPHS2* gene that results in disease only when it is associated with certain 3' alleles in *NPHS2* as a result of a dominant-negative effect [4]. As a result, the inheritance patterns in such affected families do not necessarily follow a classic recessive mode of inheritance. These unusual modes of inheritance are extremely exciting and provide novel insights into unusual genetic patterns that can result in disease manifestation. It is possible that such patterns of inheritance might prove to be the norm in disease genetics, given that there is still a vastly understood portion of causal or associated risk underlying Mendelian and complex diseases.

Founder populations

Other than the Finnish population described in Chapter 4, there are several other founder populations that can provide advantages for mapping genes and variants in Mendelian and complex diseases using whole-exome or whole-genome sequencing such as the Hutterites [5], Ashkenazi Jews [6] and French Canadians [7]. Even though there might be genes and variants that were lost as a result of bottlenecks in these populations, the unique genetic architecture of these populations can aid in the discovery of rare and low-frequency variants in Mendelian and complex diseases.

A POSTSCRIPT

It is incredibly exciting that human genetics discovery has been revolutionized by improving technologies developed through decades of research. We are now in a new era of human disease genetic mapping as a result of the wide application of whole-exome and whole-genome sequencing. With this come new challenges and opportunities for developing new methodologies and approaches to understand the coding variation in our exomes and how naturally occurring human variation can confer risk for diseases. Personally, I find it extremely challenging but exciting to be a human geneticist in this era where we can now rapidly identify and understand human variation in diseases. Hopefully these discoveries and knowledge can translate into treatment and drug discovery for various Mendelian and complex diseases at a much faster pace in the next decade.

REFERENCES

1. Ashley EA, Butte AJ, Wheeler MT, Chen R, Klein TE, et al. (2010) Clinical assessment incorporating a personal genome. Lancet 375: 1525-1535.

2. Yang Y, Muzny DM, Reid JG, Bainbridge MN, Willis A, et al. (2013) Clinical whole-exome sequencing for the diagnosis of mendelian disorders. N Engl J Med 369: 1502-1511.

3. Weedon MN, Cebola I, Patch AM, Flanagan SE, De Franco E, et al. (2014) Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. Nat Genet 46: 61-64.

4. Tory K, Menyhard DK, Woerner S, Nevo F, Gribouval O, et al. (2014) Mutation-dependent recessive inheritance of NPHS2-associated steroid-resistant nephrotic syndrome. Nat Genet.

5. Mroch A, Davis-Keppen L, Matthes C, Stein Q (2014) Identification of a founder mutation for maple syrup urine disease in hutterites. S D Med 67: 141-143.

6. Costa MD, Pereira JB, Pala M, Fernandes V, Olivieri A, et al. (2013) A substantial prehistoric European ancestry amongst Ashkenazi maternal lineages. Nat Commun 4: 2543.

7. Casals F, Hodgkinson A, Hussin J, Idaghdour Y, Bruat V, et al. (2013) Whole-exome sequencing reveals a rapid change in the frequency of rare functional variants in a founding population of humans. PLoS Genet 9: e1003815.